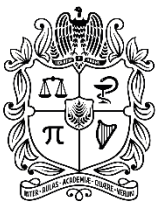


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo para la identificación de relaciones entre la información sobre los graduados de los programas de Maestría y Doctorado de la Universidad Nacional de Colombia y su tiempo de permanencia

Juan Carlos Cabrera Cárdenas

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas y Computación

Bogotá, Colombia

2017

Modelo para la identificación de relaciones entre la información sobre los graduados de los programas de Maestría y Doctorado de la Universidad Nacional de Colombia y su tiempo de permanencia

Juan Carlos Cabrera Cárdenas

Trabajo final de profundización presentado como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director (a):

Luis Fernando Niño V., PhD.

Línea de Investigación:

Sistemas Inteligentes

Grupo de Investigación:

Laboratorio de Investigación en Sistemas Inteligentes (LISI)

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas y Computación

Bogotá, Colombia

2017

Dedicatoria

A mis padres por su apoyo incondicional y por ser lo mejor que tengo en mi vida.

Agradecimientos

Agradezco al profesor Luis Fernando Niño, PhD, Profesor de la Universidad Nacional de Colombia y Director del grupo de Investigación LISI, por su paciencia, sus consejos, su guía, y sus recomendaciones, las cuáles hicieron que este trabajo llegará a buen puerto.

A los miembros del grupo de Investigación LISI, ya que siempre que se presentaba un avance del trabajo, siempre daban recomendaciones y observaciones que permitieron ir mejorando poco a poco el mismo. En especial un agradecimiento a Carlos Manuel Estevez, miembro del grupo de investigación que siempre mostró gran interés por este trabajo, y que fue un gran apoyo con sus recomendaciones y contribuciones.

Al profesor Gerardo Rodríguez Niño, PhD, profesor de la Universidad Nacional de Colombia y Director Nacional de Programas de Posgrado, por su disposición y disponibilidad a la hora tanto de brindar ideas que contribuyeron a armar esta investigación, como para facilitar la consecución de los datos, base esencial para el desarrollo de esta investigación.

También, agradezco a todas las personas que contribuyeron al desarrollo de este trabajo de investigación, sobre todo a mis compañeras y compañeros de la Dirección Nacional de Programas de Posgrado ya que me motivaron y me dieron buenos consejos.

A la Universidad Nacional de Colombia y a los profesores que guiaron con sus recomendaciones el desarrollo de la investigación y la elaboración del documento final.

A Dios que me dio la posibilidad de desarrollar esta investigación y conseguir un logro más en mi vida, y a mi familia quienes siempre me han apoyado, gracias totales.

Resumen

Durante esta investigación se llevó a cabo un proceso de descubrimiento de conocimiento (KDD por sus siglas en inglés) aplicado a datos educativos, más específicamente, a datos académicos y no académicos de los graduados de programas de posgrado (maestrías, especializaciones y doctorados) de todas las sedes la Universidad Nacional de Colombia. Se hizo énfasis en la etapa de minería de datos, teniendo en cuenta los siguientes objetivos principales, el primero buscar relaciones entre estos datos académicos y no académicos de los graduados con respecto al tiempo que les tomó graduarse, y el segundo desarrollar un modelo que pudiera dar una probabilidad de lo que puede suceder con los estudiantes que hasta ahora están ingresando a estos programas de posgrado.

Finalmente, se implementó una visualización de los resultados, tanto de los patrones relacionales como de las predicciones, en el Sistema de Autoevaluación de los Programas de Posgrado para que los programas cuando estén planteando su plan de mejoramiento puedan observar y tomar decisiones de acuerdo a la situación de cada uno de ellos.

Palabras clave: Reglas de Asociación, Minería de Datos, Tiempo de permanencia, Educación y KDD.

Abstract

In this research, a knowledge discovery process (KDD) was applied to educational data, more specifically, to academic and non-academic data of graduates of Universidad Nacional de Colombia. Focus was put on the data mining stage, keeping into account the following main goals, the first one was to look for relationships between academic and non-academic data of the graduates with respect to the time it took them to graduate, and the second one was to develop a model that could give a probability of what could happen with the students who were just admitted to the graduate programs.

Finally, a visualization of the results, both of the relational patterns and of the predictions, was implemented in the Auto-Evaluation System of the Graduate Programs at the Universidad Nacional de Colombia so that the programs could use it when they are preparing their improvement plans.

Keywords: Association rules, Data Mining, Graduation Time, Education and KDD

Contenido

Pág.

Resumen ..... **IX**

Lista de figuras ..... **XIII**

Lista de tablas ..... **XV**

Introducción **1**

1. Marco teórico **3**

1.1 Descubrimiento de Conocimiento en Bases de Datos KDD 3

1.1.1 Estudio del Problema 6

1.1.2 Selección de la información objetivo 6

1.1.3 Limpieza de Datos 6

1.1.4 Reducción y transformación de datos 7

1.1.5 Hacer coincidir los objetivos del problema con el KDD 8

1.1.6 Exploración de análisis y selección de hipótesis 8

1.1.7 Aplicación de métodos de minería de datos 8

- Generación de reglas de asociación 9
- Arboles de Decisión 11
- Clasificadores Bayesianos Ingenuos (*Naive Bayes*) 13
- Bosques aleatorios (*Random Forest*) 14

1.1.8 Evaluación de patrones 15

- Métricas de Rendimiento 15
- Curvas ROC 16

1.1.9 Presentación del conocimiento 17

1.2 Descubrimiento de conocimiento en educación 18

2. Identificación del problema y datos 23

2.1 Planteamiento del problema 23

2.2 Objetivos 24

2.2.1 Objetivo General 24

2.2.2 Objetivos Específicos 25

2.3 Metodología 25

2.4 Análisis Descriptivo 26

3. Análisis de asociación y generación del modelo 43

3.1 Preprocesamiento 43

3.2 Generación de clasificación de registros 47

3.3 Reglas de Asociación 51

3.4 Generación del Modelo 58

3.5 Evaluación del Modelo 62

3.5.1 Matrices de confusión 63

3.5.2 Curvas ROC 66

4. Predicción de casos nuevos y adecuación en el Sistema de Autoevaluación 73

4.1 Predicción de casos nuevos 73

4.2 Módulo en el Sistema de Autoevaluación 81

5. Conclusiones y recomendaciones 87

5.1 Conclusiones 87

5.2 Trabajo Futuro 89

Bibliografía 91

Lista de figuras

Pág.

Figura 1-1: Fases del proceso de KDD 5
Figura 1-2: Estructura de un árbol de decisión 12
Figura 1-3: Diagrama de Flujo para el algoritmo *Naive Bayes* 14
Figura 1-4: Ejemplos de curvas ROC con diferente área bajo la curva 17
Figura 2-1: Número de programas por sede y por facultad, clasificados por nivel 27
Figura 2-2: Diagrama de estados de un estudiante de posgrado en la Universidad Nacional 29
Figura 2-3: Distribución de graduados por Sede y Facultad 30
Figura 2-4: Número de graduados por sexo 31
Figura 2-5: Clasificación de la edad de los graduados por nivel académico de posgrado 32
Figura 2-6: Atributos Estado Civil, Estrato y Discapacidad para los graduados 33
Figura 2-7: Área de conocimiento del programa del que se graduó cada estudiante 34
Figura 2-8: Distribución de los graduados por tipo de plan de estudios y nivel 35
Figura 2-9: Mapa del árbol de la variable Ciudad de Nacimiento 36
Figura 2-10: Tipo de convenio dividido por sede y nivel académico de posgrado (Sede Bogotá) 37
Figura 2-11: Tipo de convenio dividido por sede y nivel académico de posgrado (Sede Amazonía, Sede Caribe y Sede Palmira) 37
Figura 2-12: Tipo de convenio dividido por sede y nivel académico de posgrado (Sede Manizales y Sede Medellín) 38
Figura 2-13: Distribución de los graduados por Núcleo Básico del Conocimiento (Nivel de Doctorado) 39
Figura 2-14: Distribución de los graduados por Núcleo Básico del Conocimiento (Nivel de Maestría) 40
Figura 2-15: Distribución de los graduados por Núcleo Básico del Conocimiento (Nivel de Especialización) 41
Figura 3-1: Distribución de las etiquetas de clasificación en el conjunto de datos 49
Figura 3-2: Distribución de las etiquetas de clasificación a nivel de Especialización 50
Figura 3-3: Distribución de las etiquetas de clasificación a nivel de Maestría 50
Figura 3-4: Distribución de las etiquetas de clasificación a nivel de Doctorado 51
Figura 3-5: Diagrama para obtener reglas de asociación en *Orange* 52
Figura 3-6: Diagrama del modelo de clasificación en Orange 59
Figura 3-7: Ejemplo de visualización de un árbol de decisión 60
Figura 3-8: Visualización de los arboles obtenidos del *Random Forest* 61

Figura 3-9: Visualización de los resultados del algoritmo Naive Bayes por medio de un Nomograma 62
Figura 3-10: Matriz de confusión para el algoritmo *Naive Bayes* 63
Figura 3-11: Matriz de confusión para el algoritmo *Random Forest* 64
Figura 3-12: Matriz de confusión para el algoritmo de árbol de decisión 65
Figura 3-13: Curva ROC para la clase "Entre duración de programa y el doble de duración de programa" 67
Figura 3-14: Curva ROC para la clase "Igual a duración de programa" 68
Figura 3-15: Curva ROC para la clase "Mayor al doble de duración de programa" 69
Figura 3-16: curva ROC para la clase "Menor a duración de programa" 70
Figura 3-17: Medidas de rendimiento para los algoritmos de clasificación utilizados en el modelo 71
Figura 4-1: Distribución de los datos de los admitidos divididos por sede y nivel 74
Figura 4-2: Distribución de los admitidos por sede y nivel, clasificándolo por el uso de convenios 75
Figura 4-3: Proporción de convenios para los admitidos de la sede Bogotá 76
Figura 4-4: Distribución de los admitidos a los programas de posgrado por su área de conocimiento 77
Figura 4-5: Diagrama en Orange para hacer la predicción de los admitidos 78
Figura 4-6: Predicción de los admitidos por nivel de programa (*Naive Bayes*) 79
Figura 4-7: Predicción de los admitidos por nivel de programa (*Random Forest*) 80
Figura 4-8: Predicción de los admitidos por nivel de programa (Árbol de Decisión) 81
Figura 4-9: Módulos principales del Sistema de Autoevaluación de Posgrados 82
Figura 4-10: Opciones disponibles en el módulo de Plan de Mejoramiento 83
Figura 4-11: Visualizaciones disponibles en la opción "Plan de mejoramiento" 83
Figura 4-12: Visualización de la opción análisis de permanencia 84
Figura 4-13: Presentación de Nomogramas en el SAPUN para cada Facultad 84
Figura 4-14: Presentación de los datos de predicción y los archivos fuente de la investigación 85

Lista de tablas

Pág.

Tabla 1-1: Ejemplo de una transacción de canasta de mercado (*market basket transaction*) 9
Tabla 1-2: Matriz de confusión para un problema de clasificación binaria 15
Tabla 1-3: Matriz de confusión para un problema de clasificación SI / NO 15
Tabla 3-1: Variables que caracterizan al programa 45
Tabla 3-2: Variables no académicas que describen a cada graduado 45
Tabla 3-3: Variables académicas que caracterizan a cada graduado 46

Introducción

Cuando se habla de información en el mundo de hoy, se tiene que hablar de que cada segundo en el mundo se están generando al mismo tiempo imágenes, videos, *tweets*, *blogs*, entre otros datos, sobre cualquier tema, evento o suceso que pueda ocurrir, y esta situación ocurre en la educación y en particular de la Universidad Nacional de Colombia, en donde se generan semestralmente cantidades muy grandes de información: matriculados, graduados, admitidos, notas y publicaciones de profesores, entre muchas otras. Y es ahí en donde surge la necesidad de empezar a tomar toda esa cantidad de información y poder "hacer algo" con ella, y para esto se utilizan ramas de la computación que precisamente están enfocadas a hacer análisis de grandes cantidades de información y poder generar un conocimiento que le ayude a las personas directamente implicadas con esta información, a mejorar sus procesos de toma de decisiones.

Ahora bien, cuando se habla de un proceso de descubrimiento de conocimiento en bases de datos educacionales, se puede observar cómo nace el concepto de minería de datos Educacional [1], y a partir de ahí muchos trabajos se desprenden de ese concepto [2], en los cuales su único fin es ayudar a los profesores o a las mismas instituciones a tener una vista más completa y poder tomar decisiones basados en el análisis e hipótesis realizados por cada autor. De estos trabajos se puede observar que hay bastantes que se centran en hacer predicción basándose en un modelo, y esa predicción es la que al final le va a servir a cada institución para poder tomar medidas para solucionar la problemática.

En este trabajo se llevó a cabo un proceso de Descubrimiento de conocimiento en Bases de Datos (KDD, por sus siglas en inglés) aplicado precisamente a datos educativos. Para esto se tomaron datos de los graduados de la Universidad Nacional de Colombia que han tenido los programas de Especialización, de Maestría y de Doctorado, desde el año 2009 hasta el último semestre de 2016, y gracias a la Dirección Nacional de Programas de Posgrado, su Sistema de Autoevaluación y las diferentes dependencias de la Universidad Nacional, fue posible obtener variables académicas y no

académicas complementarias de esos graduados, quedando así con casi 20.000 registros para el análisis.

Una vez se tenían estos datos se procedió a realizar la clasificación de cada registro dependiendo del tiempo que duraron en la Universidad hasta graduarse, esto debido a que una de las mayores problemáticas que tienen los programas de posgrado en la Universidad Nacional, es que hay muchos estudiantes que por diferentes razones no terminan sus estudios, y de los pocos estudiantes que si los terminan, la mayoría no lo hacen en el tiempo que cada programa ha estipulado para tal fin. Lo anterior representa un problema no solo para el programa, sino para el estudiante que ve sus metas truncadas, o que debe gastar mucho más dinero del presupuestado inicialmente para obtener su título.

Por tal motivo este trabajo tienen como fin primordial ayudar a que los programas puedan establecer estrategias de choque dentro de sus planes de mejoramiento, no solo académicas, sino también de bienestar, que ayuden a los estudiantes que hasta ahora están ingresando al programa para que ellos puedan terminar sus estudios en el tiempo que cada programa ha estipulado. Para este fin se habilitó un módulo en el Sistema de Autoevaluación de Posgrados, en donde se presenta, en primer lugar, las relaciones de asociación más importantes que se pudieron obtener del modelo con los datos de las personas ya graduadas, y, en segundo lugar, el listado de las personas nuevas con la respectiva predicción obtenida del modelo anterior y que en teoría podría ayudar a los programas a disminuir el número de personas que se gradúan después del periodo estipulado por el programa, incluso algunos gastando hasta más del doble del tiempo.

1. Marco teórico

Actualmente, las organizaciones han entendido la necesidad y la posibilidad que les puede brindar sacarle provecho al gran volumen de información que se genera todos los días, no solo para tomar mejores decisiones, sino también para aumentar sus ganancias. El descubrimiento de conocimiento en bases de datos (KDD, del inglés *Knowledge Discovery in Databases*) es un proceso iterativo, no trivial, dentro del cual se quiere buscar patrones válidos para el análisis del usuario final, es decir, es un proceso que combina dos conceptos importantes, análisis y conocimiento [3] [4].

En este capítulo se hará un recorrido por las etapas de un proceso de KDD, los algoritmos de clasificación y búsqueda de patrones más utilizados, y las aplicaciones del KDD, sobre todo en el campo de la Educación.

1.1 Descubrimiento de Conocimiento en Bases de Datos KDD

Históricamente al proceso de extracción de conocimiento en una base de datos se le ha llamado de diferentes maneras, minería de datos, extracción de conocimiento, descubrimiento de información, arqueología de información, procesamiento de patrones de información, entre varios nombres más. Pero no fue sino hasta 1989, cuando Piatetsky-Shapiro utilizó por primera vez el concepto de extracción del conocimiento en bases de datos o KDD (*Knowledge Discovery in Databases*), y allí utilizó este término para decir que el conocimiento es el fin del producto de un descubrimiento de información dada, y para conseguirlo es necesario hacer uso de técnicas de inteligencia artificial, de aprendizaje de máquina, de estadística y de bases de datos [5].

Un proceso de KDD se puede definir como un proceso no trivial para identificar patrones válidos, nuevos, potencialmente útiles, y, en últimas entendibles, presentes en un conjunto de datos [6] , y es allí donde surge la principal diferencia de conceptos entre la minería de datos y el KDD. Un

proceso de KDD se realiza en varias fases que pueden lograr ese objetivo general de obtener conocimiento no trivial de un conjunto de datos, mientras que la minería de datos es una de las fases del KDD. Sin embargo, se puede decir que la minería de datos es una de las fases más importantes, por no decir la más importante, en el proceso de KDD, y es por eso que muchos autores atribuyen el concepto de minería de datos a todo el proceso completo [6].

Los procesos de KDD son tan importantes hoy en día que tienen muchos campos de aplicación, por ejemplo, en cualquier negocio es muy común que se utilice por lo menos la fase de minería de datos para poder analizar el mercado, la situación financiera, los probables escenarios futuros y análisis de clientes, entre otras aplicaciones [7] . En la ciencia, los procesos de KDD son muy usados en áreas como la astronomía, biología molecular y geología, entre otros. También es muy útil usar los procesos de KDD para el cuidado y comportamiento en términos de salud [8], detección de fraudes, análisis de deportes y para el análisis de datos en el ámbito educativo [9]. Además, con el crecimiento de la capacidad de almacenamiento de datos en las últimas décadas, y con el crecimiento de las redes sociales, allí también tienen mucha aplicación los procesos de KDD; por ejemplo, *Facebook*, *Twitter* o *Google* utilizan la información que le provee cada usuario diariamente, tanto para recomendar nuevos contenidos, como para dar a organizaciones o empresas información de los usuarios que puedan ser de su interés [6][10][11].

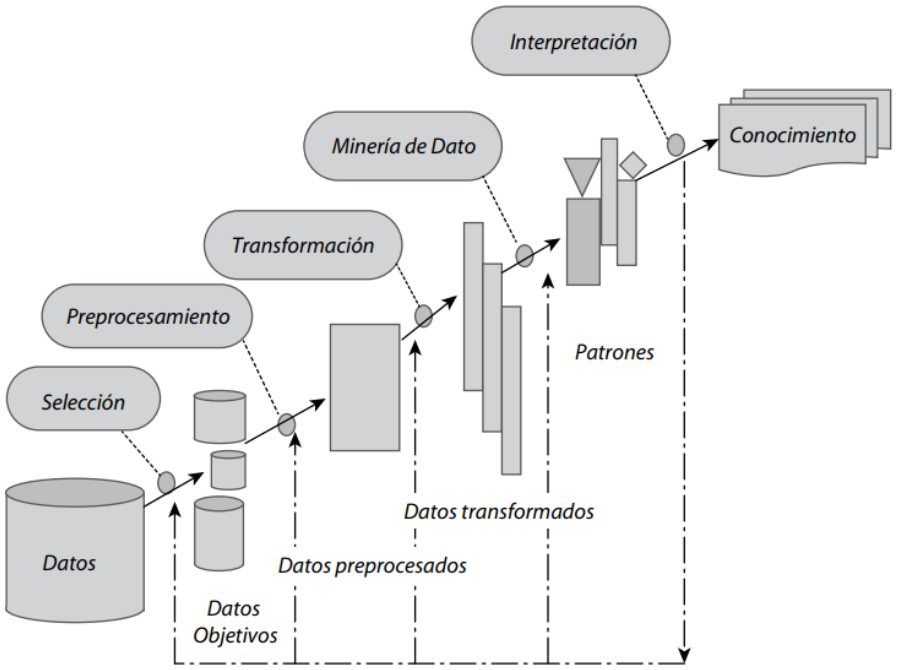
Un proceso de KDD, como ya se ha mencionado anteriormente, se encuentra conformado por una serie de pasos o fases las cuáles hacen que se pueda cumplir el objetivo primario, que es generar nuevo conocimiento a partir de los datos disponibles. Un proceso de KDD es un proceso cíclico que incluye etapas que se realizan de manera secuencial, es decir si por algún motivo al final del proceso, no se obtienen los resultados esperados, es posible hacer de nuevo todas las fases cambiando algo de lo que se haya hecho en alguna de ellas para tratar de obtener un mejor resultado, y es secuencial debido a que la salida de cada una de las fases es la entrada para la siguiente. Las 9 fases de un proceso de KDD son:

1. Estudio del problema.
2. Selección de la información objetivo.
3. Limpieza de datos.
4. Reducción y transformación de datos.

5. Hacer coincidir los objetivos del problema con el KDD.
6. Exploración de análisis y selección de hipótesis.
7. Aplicación de métodos de minería de datos.
8. Evaluación de patrones.
9. Presentación del conocimiento. [5][12]

En la Figura 1-1 se ilustran las fases de un proceso de KDD y a continuación se da una explicación breve de cada una de estas fases, describiendo en qué consisten, y mencionando los métodos y algoritmos que se usan para llevarlas a cabo:

Figura 1-1: Fases del proceso de KDD



Tomado de [3]

1.1.1 Estudio del Problema

En esta fase lo que se hace es desarrollar una comprensión del dominio de la aplicación, entender de qué se trata el problema y a dónde se quiere llegar, identificar un objetivo de un proceso de KDD desde el punto de vista del cliente o usuario final [5].

1.1.2 Selección de la información objetivo

El siguiente paso de un proceso de KDD es seleccionar el conjunto de datos o subconjunto de datos que mejor se adapte para obtener una solución satisfactoria del problema. En esta fase no solo se incluye la selección de los datos, sino también la obtención de los mismos. Vale anotar que se hace mención de este paso porque en algunos casos, puede diferir bastante el conjunto de datos que serían necesarios y el conjunto de datos que están disponibles para llevar a cabo el análisis [5] [12]. En este paso es importante poder incluir, la mayor cantidad de variables disponibles, por ejemplo, si se quiere hacer un análisis en el sector educativo es importante incluir no solo variables académicas, sino variables que describan a cada individuo, como lo pueden ser variables sociales, económicas o que tengan que ver con su formación educativa, entre otras, vale anotar que la selección de las variables sólo se puede hacer una vez se tenga claro el problema que se quiere abordar.

1.1.3 Limpieza de Datos

En esta tercera etapa, lo que comúnmente se lleva a cabo es una limpieza de datos, que consta de hacer una revisión del conjunto de datos obtenido en la fase anterior para observar y tratar de eliminar datos con ruido (modificación de valores originales) [13], *outliers* (datos con características que son considerablemente diferentes, comparados con la mayoría de los otros datos dentro del conjunto de datos) [13] e información irrelevante [14].

Para hacer esta limpieza, en primer lugar, se debe hacer un trabajo de tratar de completar al máximo los datos faltantes acudiendo de nuevo a la fuente del conjunto de datos, en el punto en el que ya no se puedan completar más; entonces existen diferentes técnicas que permiten que éste proceso no se haga de manera manual, por ejemplo, si existen pocos datos perdidos, una solución puede ser no tener en cuenta esos registros que les falta información. Sin embargo, puede ocurrir que la cantidad de datos faltantes sea alta, entonces en tal caso la solución difiere, ya que se puede hacer uso de algunas técnicas y mediciones estadísticas (valor esperado, media, mediana, muestreo) para completar los datos faltantes [15].

1.1.4 Reducción y transformación de datos

En esta fase lo que se hace con los datos es reducir el número de variables, es decir, escoger las variables que serán el objeto del análisis, y con las cuales se puede hacer un proceso de extracción del conocimiento más efectivo. También, además de escoger las variables que se tendrán en cuenta, se debe hacer un análisis a cada variable de tal modo que queden transformadas para así poder aplicar los métodos de minería de datos. Para hacer esta transformación de datos se pueden aplicar varias técnicas como lo son la agregación, la normalización o la discretización [15]. Con estas técnicas lo que se quiere hacer es reducir la redundancia de variables y convertir los datos que puedan estar muy dispersos, de manera que los algoritmos de minería puedan tener un tiempo de ejecución óptimo, y que la calidad del modelo de conocimiento sea el mejor [12].

También es muy común que este proceso de reducción de dimensiones se haga utilizando la tabla de datos y empezar a reducir horizontalmente las tuplas repetidas y, en el caso en que la reducción se haga verticalmente, se consigue empezar a eliminar variables que no aportan nada al modelo, como lo pueden ser, por ejemplo, identificadores, llaves foráneas, o campos que dependen de otros campos, como la edad o la fecha de nacimiento [3].

1.1.5 Hacer coincidir los objetivos del problema con el KDD

Una vez se tiene el conjunto de datos pre-procesado, en el proceso de KDD se procede a retomar el objetivo del problema planteado en el paso 1 y a hacer un análisis de cuál sería el proceso que mejor se acomode tanto al conjunto de datos como al objetivo del KDD, si se trata de una clasificación [16] [17] [18], agrupación (*clustering*) [19] [20] o una asociación [5] [8].

1.1.6 Exploración de análisis y selección de hipótesis

En esta fase, una vez se tienen claros los objetivos que se quieren cumplir y los datos ya limpios y pre-procesados, se deben empezar a visualizar los diferentes algoritmos de minería de datos; esto con el fin de escoger el algoritmo que mejor se adecue tanto a los datos, como a los objetivos propuestos [5] [8].

1.1.7 Aplicación de métodos de minería de datos

Esta fase de minería de datos es la más importante dentro del proceso de KDD, el objetivo principal es obtener los patrones relacionales; para eso también se debe tener en cuenta las fases anteriores, y de esa manera saber cuál será el algoritmo que se va a usar para cumplir con el objetivo. Además es importante saber si el objetivo que se quiere cumplir requiere que se genere un modelo de tipo predictivo o descriptivo, en el primero se obtiene un modelo el cual es capaz de estimar valores futuros o desconocidos a partir de variables independientes o predictivas, por ejemplo, predecir si nuevos clientes son potencialmente buenos, es decir, que van a hacer grandes compras o al menos que sus compras serán regulares en el tiempo, todo eso basado en su estado civil, edad, género y profesión. En los modelos descriptivos no se predice ninguna variable, lo que se hace es identificar patrones que puedan explicar un acontecimiento, agrupan y resumen los datos para, por ejemplo, identificar grupos de personas con gustos similares o identificar patrones de compra en una tienda comercial [3].

Entonces, aparte de escoger el algoritmo más apropiado, también se debe tener en cuenta cuales son el tipo de modelo y los parámetros más apropiados, siempre pensando en lograr cumplir con el objetivo planteado en el paso 1.

Para realizar la fase de minería de datos existen muchos algoritmos, cada uno con su características y utilidad particular, los algoritmos más usados son: K-vecinos más cercanos (KNN) [21] , redes neuronales [21], árboles de decisión [21] [22], máquinas de vectores de soporte (SVM) [21] , el algoritmo a priori [23] y los clasificadores bayesianos ingenuos [24], entre otros. A continuación se profundizará en algunos de ellos que son de utilidad para la presente investigación:

▪ Generación de reglas de asociación

La generación de reglas de asociación es un problema muy común y que últimamente ha tomado mucha fuerza, ya que da la posibilidad de descubrir relaciones escondidas dentro de los datos. En un principio fue desarrollada para analizar transacciones de canasta de supermercado (*market basket transactions*), estas transacciones están compuestas de un identificador del cliente, y de una lista de objetos que cada uno de los clientes compró en un supermercado (ver Tabla 1-1) [25].

Tabla 1-1: Ejemplo de una transacción de canasta de mercado (*market basket transaction*)

TID	**Ítems**
1	{Pan, Leche}
2	{Pan, Pañales, Cerveza, Huevos}
3	{Leche, Pañales, Cerveza, Gaseosa}
4	{Pan, Leche, Pañales, Cerveza}
5	{Pan, Leche, Pañales, Gaseosa}

Tomado de [25]

Además de su aplicación en el análisis de compras de supermercado, la búsqueda de relaciones también ha tenido aplicación en otros campos como lo son la bioinformática, diagnóstico médico, minería web y análisis de información científica, entre muchas otras aplicaciones. La búsqueda de relaciones de asociación para el conjunto de ítems comprado en un supermercado, es un problema motivado por la necesidad de guardar información acerca de la compra de ítems efectuada en cada

transacción; el objetivo principal era ser capaces de encontrar reglas como “Un Cliente que compró productos de tipo ***X*** y ***Y***, también comprará productos ***Z*** y ***W*** con una probabilidad p” [25] [26].

Las reglas de asociación son expresiones de implicación que tienen la forma $X -> Y$ en donde X y Y son conjuntos de ítems que no tienen elementos en común, y que deben contar con una confianza y un soporte, obtenido de los mismos datos. El soporte se define como la proporción de transacciones que contiene un conjunto de ítems dentro de todo el conjunto de datos, y la confianza determina el número de transacciones en las que aparece el conjunto Y con respecto al total de transacciones en el que aparece el conjunto de ítems X . El soporte es una medida primordial dentro de la evaluación de las reglas obtenidas en un modelo, ya que las reglas que tengan un soporte muy bajo, se pueden eliminar porque son reglas no interesantes. Y aunque también se busca que las reglas encontradas tengan alta confianza, obtener esa medida ideal no implica causalidad, es decir, aunque una regla tenga confianza alta y soporte alto, también es posible que esa regla no siempre se cumpla [25].

En todo caso, la búsqueda de todas las reglas posibles y el cálculo del soporte y la confianza de cada una de ellas es un procedimiento extremadamente demorado y complejo, entonces lo que usualmente hacen los algoritmos es empezar a eliminar las reglas que no pasen unos límites mínimos tanto de confianza como de soporte y, de esta manera, se puede ahorrar el cálculo de operaciones como, por ejemplo, la confianza de las reglas que no cumplan con el soporte mínimo. Dentro de esta estrategia adoptada por la mayoría de algoritmos que generan reglas de asociación se pueden definir dos subtareas:

- **Generación de conjuntos de ítems frecuentes:** El objetivo es encontrar todos los conjuntos de datos que cumplan con el límite del soporte mínimo. El costo computacional de hacer el cálculo del soporte con una aproximación de fuerza bruta es muy alto; por esto es más recomendable usar una aproximación como la que utiliza el algoritmo A priori. Esta aproximación lo que dice es que se debe ir calculando el soporte para los conjuntos de un solo ítem, y desde ese momento ya se puede eliminar los conjuntos de ítems que no cumplan con el soporte mínimo, ya que como dice el principio A priori: “Si un conjunto de

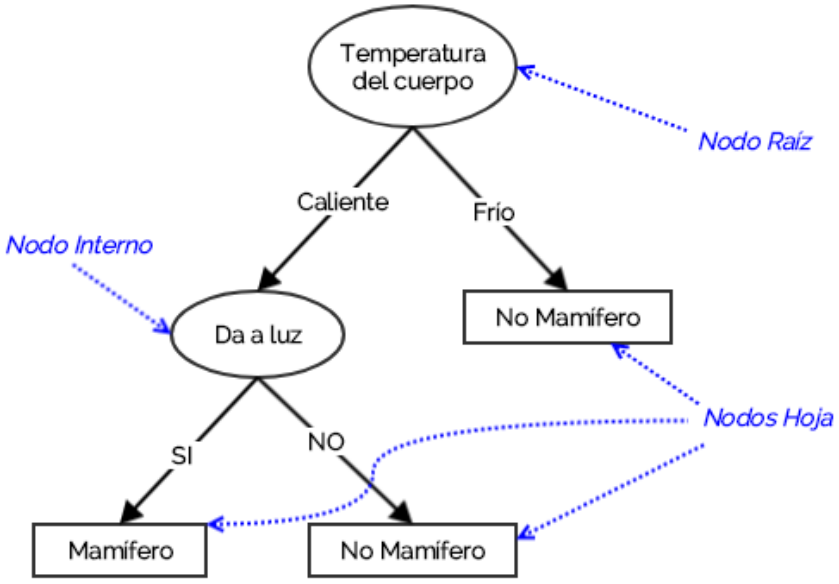
ítems es frecuente, entonces todos sus subconjuntos también deben ser frecuentes". Una vez se haga ese primer barrido, se procede a calcular el soporte para los conjuntos de 2 ítems, pero solo teniendo en cuenta los ítems que se obtuvieron del paso anterior, y nuevamente se debe hacer el filtro de solo mantener los conjuntos de datos que tengan un soporte mínimo, este proceso iterativo se mantiene hasta que no se tenga posibilidad de hacer más combinaciones con los conjuntos de ítems resultantes en cada iteración [25].

- **Generación de reglas:** La obtención de las reglas de asociación también tienen el mismo problema de los conjuntos de ítems frecuentes, esto debido a que para cada uno de los conjuntos de ítems frecuentes obtenidos se deben obtener subconjuntos no vacíos que permitan obtener reglas a partir de las diferentes combinaciones, y cada una de ellas debe pasar por un límite de confianza mínima. Lo que en este caso hacen los algoritmos como el Apriori es ir revisando cada regla y si alguna no cumple con la confianza mínima, entonces las reglas que se puedan obtener del subconjunto de cada causa, tampoco tendrán una confianza igual o mayor al límite, por lo cual también deberían ser eliminadas. De este modo, el procedimiento no es computacionalmente costoso, y se pueden obtener las reglas de asociación verdaderamente importantes para el problema [25].

▪ Arboles de Decisión

Es un método de clasificación cuyo fin es representar por medio de nodos y arcos la relación que exista entre los diferentes atributos de los datos, los nodos internos representan cada uno de los atributos, y los arcos los posibles valores que pueda tener ese atributo (Ver Figura 1-2). Cada uno de los caminos o las ramas de los árboles de decisión terminan en un nodo que no tendrá arcos o nodos hijos, este nodo es llamado hoja y cada uno de ellos está en representación de cada clase o etiqueta que produjo el modelo, o que ya tienen asignados los datos. Para hacer la clasificación de un registro simplemente se debe comenzar en la raíz de árbol y seguir los arcos según el valor que tenga el registro para cada uno de los atributos, al final la etiqueta del nodo hoja al que se llegue es la clasificación de ese registro [27].

Figura 1-2: Estructura de un árbol de decisión



Tomado de [27]

En este trabajo se usó una implementación de los árboles de decisión llamada algoritmo C4.5; este algoritmo fue desarrollado en 1993 por JR Quinlan, como una extensión o mejora del algoritmo ID3 que él mismo desarrolló en 1986 [28] , y que se basa en el algoritmo de Hunt, las características más importantes de este algoritmo son (ver [27]):

- Se consideran todas las pruebas que sean posibles para dividir el conjunto de datos y se selecciona el atributo que mayor ganancia de información le pueda generar.
- Una de las características más importantes de este algoritmo es que funciona bien tanto con variables discretas como continuas.
- Si todos los registros del conjunto de datos de entrenamiento llegan a una clase Y, entonces el nodo hoja tendrá la etiqueta Y.
- Si los registros son de clases diferentes, entonces el algoritmo selecciona el atributo que le permita dividir el conjunto en subconjuntos más pequeños, y de este modo se sigue aplicando recursivamente el algoritmo a los nodos hijos. Para hacer la selección, el algoritmo usa una función de ganancia de información *(Gain Ratio),* aunque también puede utilizar otro tipo de medidas como lo son la entropía, Gini o un error de clasificación [27].

El algoritmo C4.5 también hace uso de una función de poda, la cual se encarga de limpiar un poco el árbol generado para evitar sobre-entrenamiento y para generar reglas más claras, esta poda se lleva a cabo, por ejemplo, reemplazando un subárbol con la rama más utilizada, o con un nodo hoja con la etiqueta definida por la clase que tengan la mayoría de nodos de ese subárbol [27].

▪ Clasificadores Bayesianos Ingenuos (*Naive Bayes*)

Es una técnica de clasificación probabilística basada en la aplicación del teorema de Bayes con una fuerte suposición de independencia, relaciona el conjunto de atributos de los datos con la etiqueta de clase que se le debe dar a cada registro. Para la clasificación deben tenerse muy en cuenta los atributos, y lo que hace el clasificador es tomar este problema, y transformarlo en un problema de probabilidad condicional, donde se puede expresar como la probabilidad de que un registro sea de la clase Y, dado que en el registro tiene un conjunto de atributos X; siguiendo el teorema de Bayes se puede obtener (ver [26]):

$$P(Y = y\,|X = x) = \frac{P(X = x|Y = y) * P(Y = y)}{P(X = x)}$$

Donde:

$P(Y = y\,|X = x)$ Es la probabilidad posterior de Y

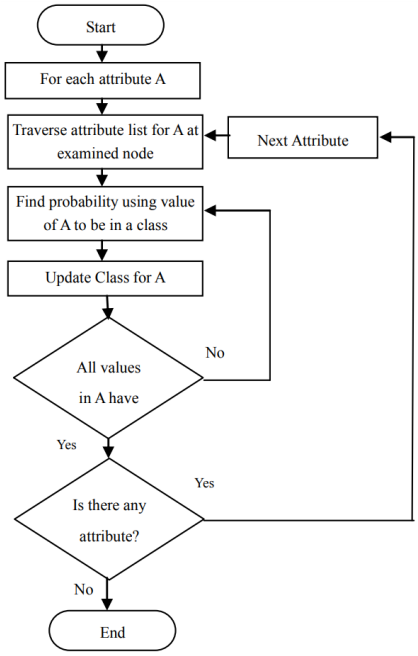
$P(X = x\,|Y = y)$ Es la probabilidad condicional de la clase

$P(Y = y)$ Es la probabilidad de que la clase está etiquetada como y

$P(X = x)$ Es la probabilidad del conjunto de atributos

Entonces para clasificar un registro lo que se busca es que el clasificador pueda maximizar la probabilidad posterior y así darle una etiqueta a este registro. El diagrama de flujo de este algoritmo es presentado en la Figura 1-3. Este algoritmo se conoce como ingenuo porque supone que todos los atributos son condicionalmente independientes, es decir, que la presencia o no de algún atributo no está relacionado con la presencia o ausencia de otro atributo. Por lo tanto, considera que cada una de las características contribuye de manera independiente a la probabilidad de que un registro pertenezca a una clase [27] [29].

Figura 1-3: Diagrama de Flujo para el algoritmo *Naive Bayes*



Tomado de [30]

▪ Bosques aleatorios (*Random Forest*)

En un método de clasificación que tiene múltiples estructuras de árbol, conteniendo múltiples clasificadores, es decir, es un conjunto de árboles simples que son capaces de determinar una etiqueta final para un conjunto de atributos dado. Contrario a los árboles de decisión explicados anteriormente, no utiliza una medida para seleccionar el mejor nodo raíz, simplemente lo va seleccionando de manera aleatoria, obteniendo así un conjunto de árboles con raíces distintas, y muy posiblemente con estructuras diferentes. Una vez obtenidos estos árboles de decisión, el algoritmo lo que hace es probar con cada uno de los registros obteniendo un etiqueta final en cada uno de los árboles, dejando como etiqueta final para el registro la que más frecuencia haya tenido dentro de todos los árboles [31] . El uso de los bosques aleatorios (*random forest*) tiene varias ventajas, lo que hace que sea uno de los métodos más usados en la actualidad:

- Produce un clasificador muy preciso.
- Funciona muy bien y de manera eficiente con conjuntos de datos de gran tamaño.

- Los bosques generados pueden ser guardados para ser usados sobre datos futuros.
- Provee estimaciones sobre cuáles variables son importantes en la clasificación [32].

1.1.8 Evaluación de patrones

▪ Métricas de Rendimiento

Una vez se utiliza alguno de los métodos y algoritmos de minería de datos, este da como salida un conjunto de patrones, que deben ser evaluados. En el caso de problemas de clasificación, es muy común hacer uso de métricas de desempeño, que son medidas que son calculadas a partir de la matriz de confusión, y ésta a su vez es obtenida a partir de la comparación entre la clase de cada registro, con la clase que le fue asignada por el algoritmo, es decir, un conteo de casos bien clasificados y mal clasificados [21] [33].

Para un problema de clasificación binaria, en la Tabla 1-2 se puede ver un ejemplo de una matriz de confusión, cada una de las entradas f_{ij} en la tabla hace referencia al número de registros de la clase i que se predijeron como clase j, luego por ejemplo, f_{01} es el número de registros que son de la clase 0, que fueron equivocadamente puestos como clase 1.

Tabla 1-2: Matriz de confusión para un problema de clasificación binaria

		Predicción	
		Clase = 1	**Clase = 0**
Clase Actual	**Clase = 1**	f_{11}	f_{10}
	Clase = 0	f_{01}	f_{00}

Tomado de [27]

Tabla 1-3: Matriz de confusión para un problema de clasificación SI / NO

		Predicción	
		SI	**NO**
Clase	**SI**	TP	FN
	NO	FP	TN

Tomado de [8]

La Tabla 1-3 representa una matriz de confusión para casos típicos SI/NO, y a partir de ella se pueden obtener las 4 categorías que se encuentran dentro de una tabla de confusión:

- TP (*True Positive*): Cantidad de registros con clase SI, y que quedaron bien clasificados
- TN (*True Negatives*): Cantidad de registros con clase NO y que quedaron bien clasificados.
- FP (*False Positives*): Cantidad de registros que fueron incorrectamente clasificados como SI
- FN (*False Negatives*): Cantidad de registros que fueron incorrectamente clasificados como NO, cuando su clase verdadera es SI.

Además, para tener una mejor vista del modelo generado, una vez se tiene la matriz de confusión, entonces se hace uso de las métricas de rendimiento, algunas de ellas son: exactitud, tasa de error, tasa de falsos positivos, sensibilidad, precisión y puntaje F, definidos así:

$$Exactitud = \frac{Número\ de\ Predicciones\ Correctas}{Total\ de\ Registros} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$Tasa\ de\ Error = \frac{Número\ de\ predicciones\ incorrectas}{Total\ de\ Registros} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$FP\ Rate\ (tasa\ de\ falsos\ positivos) = \frac{FP}{N} = \frac{FP}{(FP + TN)}$$

$$Sensibilidad\ (tasa\ de\ Verdaderos\ Positivos) = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$Precisión\ (Valor\ Predictivo\ Positivo) = \frac{TP}{TP + FP}$$

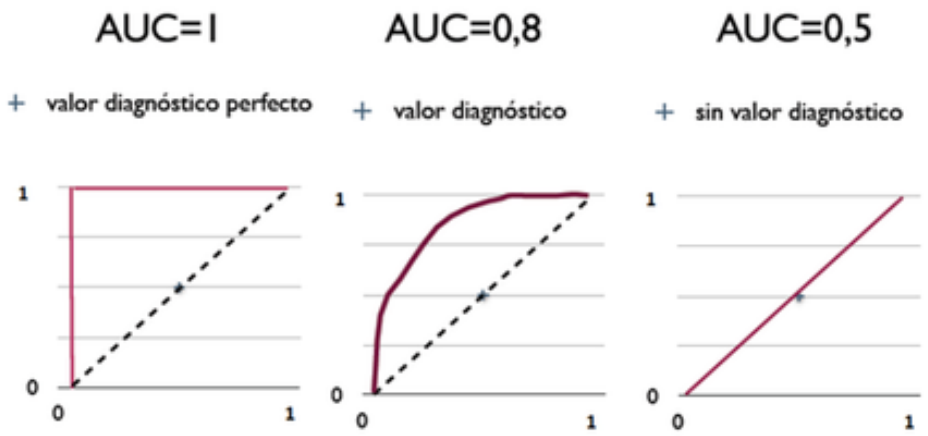
$$F\ score = Precisión * Sensibilidad$$

▪ Curvas ROC

Además de estas métricas, también se puede obtener una representación gráfica, con ayuda de la curva ROC (acrónimo de *Receiver Operating Characteristic*, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación, este tipo de gráficas son

usadas para medir el rendimiento de modelos de clasificación. Una curva ROC toma dos medidas para cada uno de sus ejes, en el eje X la tasa de falsos positivos, y en el eje Y la tasa de verdaderos positivos que mide la tasa de verdaderos positivos contra la tasa de falsos positivos, y con ella se puede obtener una métrica adicional, el área bajo la curva ROC., la cual toma valores entre 0.5 y 1, siendo 1 una medida de un clasificador ideal (ver Figura 1-4) [8] [21].

Figura 1-4: Ejemplos de curvas ROC con diferente área bajo la curva



1.1.9 Presentación del conocimiento

Una vez se tiene el modelo, los patrones totalmente identificados, y basándose en los objetivos e hipótesis planteadas antes de llevar a cabo el modelo, solo queda la posibilidad de extraer las conclusiones; observar si los patrones obtenidos con el modelo pueden dar conocimiento adicional al usuario o no, si cumplen con las expectativas que se tenían y adicionalmente observar si los resultados obtenidos pueden dar al usuario la posibilidad de implementar cambios en su modelo de negocio, mejorar su toma de decisiones, o si simplemente puede tomar estos patrones obtenidos como base para mejorar sus procesos.

1.2 Descubrimiento de conocimiento en educación

Existen muchos métodos y algoritmos que permiten extraer conocimiento de bases de datos para mejorar la toma de decisiones oportunas. Uno de los campos en el cual ha tenido aplicación este proceso es la educación, dando lugar a un nuevo concepto: minería de datos Educacional (MDE), una definición para este concepto podría ser el que da la *International Educational Data Mining Society*: "MDE es una disciplina en evolución, que tiene que ver con el desarrollo de métodos para explorar los tipos únicos de datos que provienen de ambientes educativos, y por medio de la aplicación y uso de estos métodos lograr una mejor comprensión de los estudiantes y el entorno en el que aprenden" [1].

Plantea bastantes ventajas, comparándola con métodos de investigación más conservadores como experimentos de laboratorio, estudios sociológicos o investigaciones de diseño. Sobre todo, teniendo la posibilidad en la actualidad de tener un repositorio con los diferentes datos que se pueden generar en un ámbito educativo. Lo que ahorra mucho tiempo en procesos como la consecución de los individuos, la organización de los estudios y la recopilación de datos. Luego las posibilidades que da la minería de datos de descubrir todo ese conocimiento escondido son muy amplias [12] [34].

Aunque depende en primera instancia de los datos que se van a analizar, llevar a cabo MDE, en una institución puede traer muchos beneficios: Caracterizar el comportamiento y los logros de los alumnos, mejorar el proceso de enseñanza/aprendizaje (matrículas, tutorías, evaluación, graduación), mejorar indicadores de gestión (disminuir abandono de los estudiantes, aumentar tasa de graduación, tutorías personalizadas), mejorar la infraestructura, optimizar el uso de aulas y laboratorios o mejorar la eficiencia organizacional, por solo mencionar algunos beneficios [2].

Como se mencionó anteriormente, en la minería de datos, un modelo de descubrimiento de patrones puede tener varios métodos: predicción, agrupación (*clustering*), minería de relaciones y descubrimiento de modelos, y por supuesto dentro de MDE, también son usados estos métodos.

En [2] se hace una revisión de los trabajos realizados en el marco de la MDE, haciendo una búsqueda desde el año 2005 hasta 2015, y obteniendo como resultados, 65 artículos que con algún caso de estudio utilizaron algún método o algoritmo de minería de datos para generar conocimiento. En este estudio se mostró que dentro de la literatura encontrada, en 18 ocasiones se usó clasificación, seguido por agrupación (*clustering*) con 14, asociación con 12, y predicción con 8 casos. Además de observar cuáles métodos se usaron, y cuales herramientas, también se dedicaron a hacer un análisis de los objetivos con el cual se hicieron estas investigaciones. Se obtuvo que 21 casos hicieron su investigación para predecir evaluaciones finales o rendimiento de los estudiantes, 6 para predecir el perfil y comportamiento de aprendizaje de los estudiantes y 3 para mejorar el soporte de los docentes.

Otro estudio que se realizó sobre MDE fue publicado en [35] , en el cual se presentó un análisis sobre los algoritmos y metodologías de agrupación usadas en investigaciones en los últimos 30 años, encontrando (y siguiendo la línea de que dentro de los campos que más se trabajan en la MDE están la predicción del comportamiento de los estudiantes, y hay varias investigaciones que se han enfocado sobre todo en esa nueva rama de la educación que es el e-learning o educación virtual. Dentro de este estudio también se observa que el algoritmo más usado para hacer agrupación es K-Means [36] , seguido por el agrupamiento aglomerativo, y por una variación de K-Means, Fuzzy K-means [37].

Como se puede observar, en el campo del MDE hay muchas investigaciones que se han llevado a cabo, pero siendo un poco más específico y profundo, hay algunos trabajos interesantes que tienen que ver con predicción: en [38] se hace todo un proceso de minería de datos para generar reglas de asociación obtenidas a partir de árboles de decisión, para predecir el rendimiento de estudiantes en un curso de programación; en [9] utilizaron el algoritmo J48 de árboles de decisión y una regresión lineal múltiple para predecir el comportamiento de algunos estudiantes que están en un curso virtual de Moodle; en [39] se hizo una comparación de 2 algoritmos (C4.5 y un clasificador bayesiano ingenuo), para predecir la habilidad cognitiva de algunos estudiantes, según los resultados de una prueba *on-line*; como conclusión de este estudio se observó que el algoritmo de C4.5 fue el que mejor resultado obtuvo tanto en la matriz de confusión como con las medidas de

rendimiento. Además, en [40] se llevó a cabo una comparación, en este caso basándose en datos de estudiantes de Moodle y con datos de grupos de Facebook a los cuales pertenecen estos estudiantes, tratando de predecir la personalidad de los estudiantes, y para hacer esto utilizaron algoritmos como clasificadores bayesianos, J48, *Kstar*, KNN, *OneR*, árboles de decisión, y bosques aleatorios; en esta investigación se concluyó que el mejor algoritmo de clasificación fue *OneR*, seguido por J48. Y aunque existan muchas investigaciones que también utilizan datos con orígenes educativos para hacer predicción, cada conjunto de datos funciona de manera diferente, y cada investigación tiene sus propios objetivos, por estos motivos se usan diferentes algoritmos. Por ejemplo, en [41] utilizan clasificadores bayesianos, en [42] utilizan el algoritmo *a priori* y árboles de decisión o hasta algoritmos un poco más complejos como lo son *Random Forest* [43] o el algoritmo de maximización de la esperanza (*Expectation Maximization*) [44] ; lo que se puede observar es que antes de escoger un algoritmo o método se debe hacer un buen análisis: con qué tipos de datos se cuenta y cuáles son los objetivos que se quieren cumplir al final de la investigación.

En el contexto de la Universidad Nacional de Colombia, también se han realizado investigaciones en MDE; en el año 2010, la Vicerrectoría General hizo todo un análisis tanto descriptivo como analítico sobre la permanencia estudiantil en los programas de posgrado [45]. En dicho estudio se tuvieron en cuenta no solo variables de deserción, sino también otra información que tenía que ver con las becas, y con la graduación. Cabe destacar que para este informe se tuvo en cuenta información entre 2004 y 2008. También, en 2013, se desarrolló un modelo para predecir el comportamiento académico de los estudiantes de los programas de pregrado de Ingeniería de Sistemas e Ingeniería Ambiental, para esto se tomaron datos entre 2007 y 2012, y se usaron algoritmos como clasificadores bayesianos y árboles de decisión [26].

Para este trabajo, a diferencia de lo mostrado en [26], se trabaja con tiempos de permanencia, y no con deserción, y además para el estudio se tienen en cuenta los programas de posgrado de Especialización, Maestría y Doctorado que hayan tenido graduados en la ventana de tiempo analizada Y aunque también se hace un análisis descriptivo y analítico de los tiempos de permanencia como en [45], en este trabajo se utilizan técnicas y algoritmo de minería de datos,

para generar conocimiento adicional, y se genera un módulo en el Sistema de Autoevaluación de Posgrados, para poder mostrar los resultados del trabajo a los programas académicos. Además de que la ventana de tiempo se sitúa entre 2009 y 2015, actualizando totalmente lo presentado por la Vicerrectoría, que se sitúa entre 2004 y 2008.

2. Identificación del problema y datos

En este capítulo se hace una presentación del problema que se abordó en este trabajo y los objetivos, además se presenta la metodología que se usó para llevarla a cabo. Después se hace una presentación de los datos que se utilizaron en la investigación con ayuda de un análisis descriptivo. Posteriormente, el modelo de Minería de datos que se adelantó se presenta en los capítulos 3 y 4.

2.1 Planteamiento del problema

Actualmente, en distintas áreas se ve que la generación de los datos tienen un crecimiento exponencial y esta situación no puede estar aparte de la educación, en particular, de la Universidad Nacional de Colombia, en donde se generan semestralmente grandes cantidades de información: matriculados, graduados, admitidos, notas y publicaciones de profesores, entre muchas otras. Entre estos datos y con ayuda de los sistemas de autoevaluación de la Universidad, se puede observar que una de las mayores dificultades que tienen los programas curriculares de posgrado en la Universidad Nacional de Colombia está relacionada con el tiempo de permanencia de los estudiantes para poder graduarse; el porcentaje de estudiantes que se gradúan es relativamente bajo, y entre los estudiantes que logran graduarse, un alto porcentaje no lo hacen en el tiempo que el programa ha establecido. Se presenta entonces la necesidad de hacer un análisis que permita identificar algunas causas para esta larga duración en los programas, así como las relaciones entre las características de las personas a las que les toma más tiempo graduarse con respecto a la duración del programa y las de los estudiantes que sí lo hacen en el tiempo establecido.

Con lo anterior en mente, surgió la necesidad de analizar la información que se tiene de las personas que ya se graduaron de los programas de maestría, especializaciones y doctorado de todas las sedes de la Universidad y, posteriormente, hacer una evaluación y predicción sobre los

datos de los estudiantes recién admitidos o que se encuentran iniciando algún programa de maestría o doctorado. Para llevar a cabo esto se siguieron las etapas propuestas (selección de datos, pre-procesamiento, transformación, minería de datos y evaluación e interpretación) en el proceso de KDD (*Knowledge Discovery in Databases*) [5].

Dado lo anterior, se plantean una pregunta principal para responder a través de este trabajo:

- ¿Cuáles son las causas que influyen en el tiempo de permanencia distinto al establecido para cada programa de posgrado de los estudiantes que ya obtuvieron su título?

Por consiguiente, en el presente documento se presenta el método aplicado para obtener un modelo para tratar de establecer relaciones entre la información disponible de los graduados y su tiempo de permanencia. Además, se implementó un módulo en el sistema de autoevaluación de posgrados para visualizar los resultados del modelo desarrollado. Lo anterior podrá contribuir a que los programas puedan construir sus planes de mejoramiento, usando información sobre los patrones relacionales dados por el aplicativo. De esta forma, se pueden crear estrategias tanto académicas como de bienestar, para lograr que los estudiantes se gradúen en un tiempo más cercano al tiempo de duración establecido. Lo anterior permitirá también demostrar una vez más los beneficios del uso de la MDE.

2.2 Objetivos

2.2.1 Objetivo General

Desarrollar un modelo para la identificación de relaciones entre la información sobre los graduados de los programas de Maestría y Doctorado de la Universidad Nacional de Colombia y su tiempo de permanencia.

2.2.2 Objetivos Específicos

- Identificar las variables más relevantes de los graduados de los programas de Maestría y Doctorado que permitan generar las asociaciones con respecto a su tiempo de permanencia.
- Implementar un aplicativo o módulo basado en la información suministrada por el modelo desarrollado, el cual podría ser usado por los programas curriculares de posgrado de la Universidad Nacional de Colombia para elaborar sus planes de mejoramiento.
- Aplicar el modelo desarrollado para identificar dentro de los admitidos y estudiantes que apenas empiezan su maestría y su doctorado, aquellos estudiantes que tendrían más posibilidades de tener altos tiempos de permanencia.

2.3 Metodología

Para poder llevar a cabo este proyecto final de Maestría se siguió una metodología organizada en cuatro fases, en donde cada una de ellas aporto lo necesario para cubrir los objetivos específicos anteriormente descritos. La metodología usada fue en primera instancia descriptiva, ya que se hizo análisis descriptivo inicial de los datos con los que se contó y después se generaron algunas relaciones de asociación que puedan explicar algunas causas para que los tiempos de permanencia sobrepasen los establecidos por cada uno de los programas curriculares; por lo cual la metodología se puede considerar como cualitativa. Por último, en las últimas etapas del proyecto la metodología pasa a ser experimental, ya que se debe probar el modelo predictivo que se generó a través de las diferentes actividades, y con el cual se quiere dar a conocer a cada programa curricular de posgrado las probabilidades de que los nuevos estudiantes tengan un tiempo de permanencia normal, o elevado. Para realizar cada una de las etapas, se utilizaron como base las fases de un proceso de KDD: selección de datos, preprocesamiento, transformación, minería de datos e interpretación y evaluación.

Fase 1: Selección de Datos y preprocesamiento

En esta fase se recopilaron, analizaron y prepararon todos los datos. Se realizó una limpieza de los datos, se eliminó el ruido, se trataron de completar los datos faltantes y se eliminaron todos los datos redundantes o que no aportara al modelo.

Fase 2: Transformación y Minería de Datos

Después de tener las variables seleccionadas, y los datos pre procesados, se realizó la transformación de los datos de manera que los métodos o algoritmos a utilizar posteriormente para hacer la minería de datos, no tuvieran ningún problema.

Fase 3: Construcción aplicativo

Se desarrolló un módulo integrado al SAPUN (Sistema de Autoevaluación de Programas de Posgrado de la Universidad Nacional de Colombia), que le permitirá a los programas curriculares, con base en las relaciones de asociación generadas en la fase anterior, elaborar sus planes de mejoramiento.

Fase 4: Evaluación

Una vez se produjo el modelo de generación de reglas de asociación, se hizo una fase de evaluación y pruebas. Además se generó una predicción de lo que podría llegar a pasar con los estudiantes recién admitidos a los programas de maestría, doctorado y especialización de la Universidad. Esta predicción también se muestra de manera visual en el módulo del Sistema de Autoevaluación.

2.4 Análisis Descriptivo

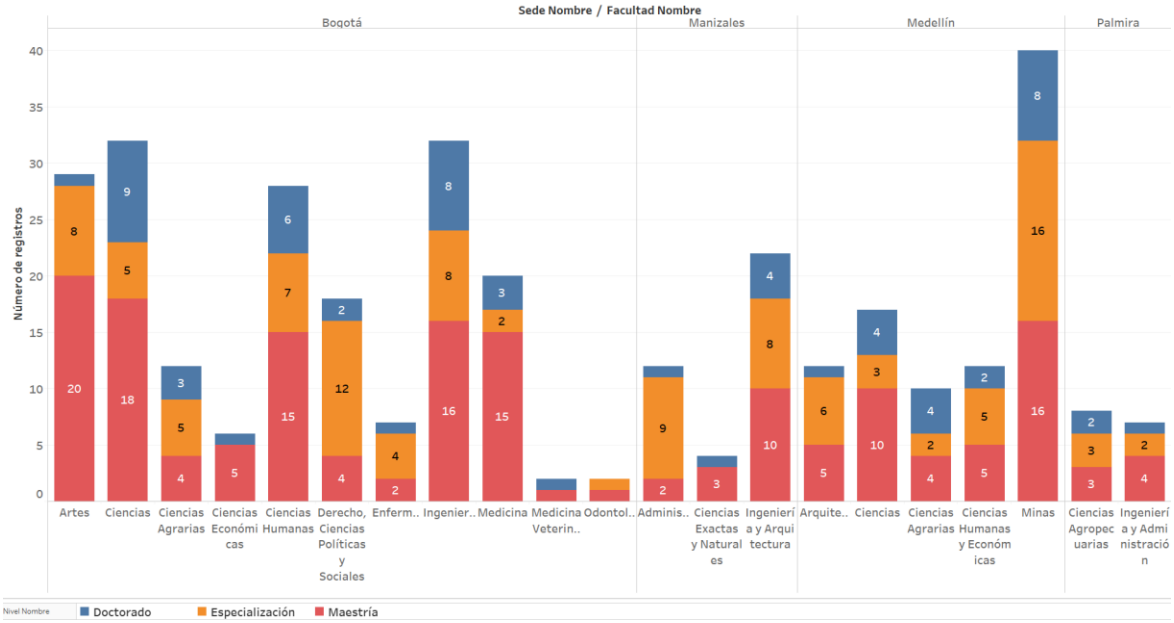
La Universidad Nacional de Colombia fue creada en 1867 por medio de la expedición de la Ley 66 del Congreso de la República, como un ente universitario con plena autonomía vinculado al Ministerio de Educación Nacional, con régimen especial, de carácter público y perteneciente al Estado.[1] Actualmente tiene 8 sedes a nivel nacional (Amazonía, Bogotá, Caribe, Manizales, Medellín, Orinoquia, Palmira y Tumaco) y próximamente abrirá otra sede en el departamento del Cesar. La Universidad Nacional se mantiene como una de las mejores universidades. no solo a nivel

[1] Naturaleza de la Universidad Nacional de Colombia, Tomado de: http://unal.edu.co/la-universidad/naturaleza.html

nacional, sino a nivel de Latinoamérica, donde se ha mantenido entre las primeras 15 universidades más importantes en los últimos años.[2]

Actualmente la Universidad Nacional de Colombia cuenta con 343 programas de posgrado activos a nivel nacional, entre doctorados (64), maestrías (163), especializaciones (76) y especialidades médicas (40) (ver Figura 2-1). Para esta investigación, en principio, se planteó realizar el análisis solamente para programas de maestría y doctorado, pero al hacer un análisis exploratorio previo, se decidió que los programas de especialización también podían hacer parte del trabajo, ya que pueden brindar, primero una mayor cantidad de datos para mejorar el análisis, y segundo tienen un comportamiento parecido a las maestrías. Algo muy diferente, en cambio, ocurre con las especialidades médicas, las cuales tienen características muy especiales, por ejemplo, la duración o el hecho de que la convocatoria sea anual.

Figura 2-1: Número de programas por sede y por facultad, clasificados por nivel

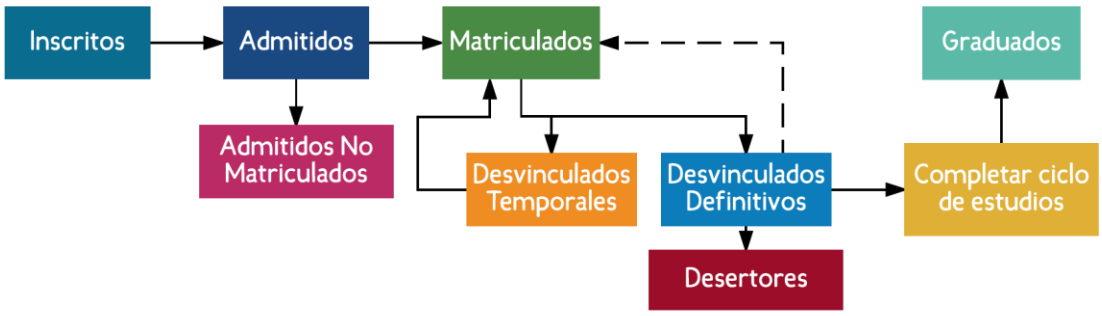


[2] *QS Latin America University Ranking* 2018, tomado de: https://www.topuniversities.com/university-rankings/latin-american-university-rankings/2018

Para el presente trabajo se tomó información sobre las personas que ingresaron y se graduaron después del 2009 y hasta el segundo semestre del 2015. Esto para poder tener los datos sin el "ruido" que podría introducir la reforma académica que se hizo en el año 2008. En primera medida, y tal como se muestra en la Figura 2-2, se puede observar el ciclo que puede tener un estudiante durante su estancia en la Universidad. En primer lugar, los estudiantes se inscriben para ser admitidos a un programa de posgrado. Después de realizar y pasar por todos los exámenes y entrevistas establecidos por cada programa para su proceso de ingreso, los estudiantes pasan a ser admitidos. En este estado, los estudiantes pueden hacer uso o no de la matrícula. Una vez están en el estado matriculados, y durante sus estudios podrán pasar a 2 estados: desvinculación temporal, la cual usualmente se presenta con los estudiantes que hacen reserva de cupo (no se matriculan por 1 o hasta 2 semestres), pero después regresan a estudiar normalmente; y el otro estado es la desvinculación definitiva, este estado se presenta por 2 situaciones, una se presenta cuando un estudiante se convierte en desertor al perder su calidad de estudiante por causas definitivas (como no aprobar 2 actividades académicas, o tener el P.A.P.A. , Promedio Aritmético Ponderado Acumulado, menor a 3.5), la otra situación se presenta cuando un estudiante logra aprobar todas las asignaturas y actividades académicas que cada programa tiene como requisitos de grado en sus planes de estudio, en ese momento los estudiantes pasan a un estado de completar ciclo de estudios, y solo les falta algún requisito o trámites para ser graduados, en el caso de las maestrías sucede que cuando un estudiante entrega su tesis o trabajo final a la facultad, éste debe pasar por una sustentación pública o una revisión de jurados, para así poder completar el ciclo de estudios, pero para todo este proceso, la facultad tiene como plazo hasta el último día de clase del semestre siguiente a la entrega, para el caso de los doctorados el requisito que usualmente trae demoras para la gradación es tener un artículo publicado en una revista especializada.[3]

[3] Artículo 19 del Acuerdo 033 de 2008 del Consejo Superior Universitario, Tomado de: http://www.legal.unal.edu.co/sisjurun/normas/Norma1.jsp?i=34725

Figura 2-2: Diagrama de estados de un estudiante de posgrado en la Universidad Nacional



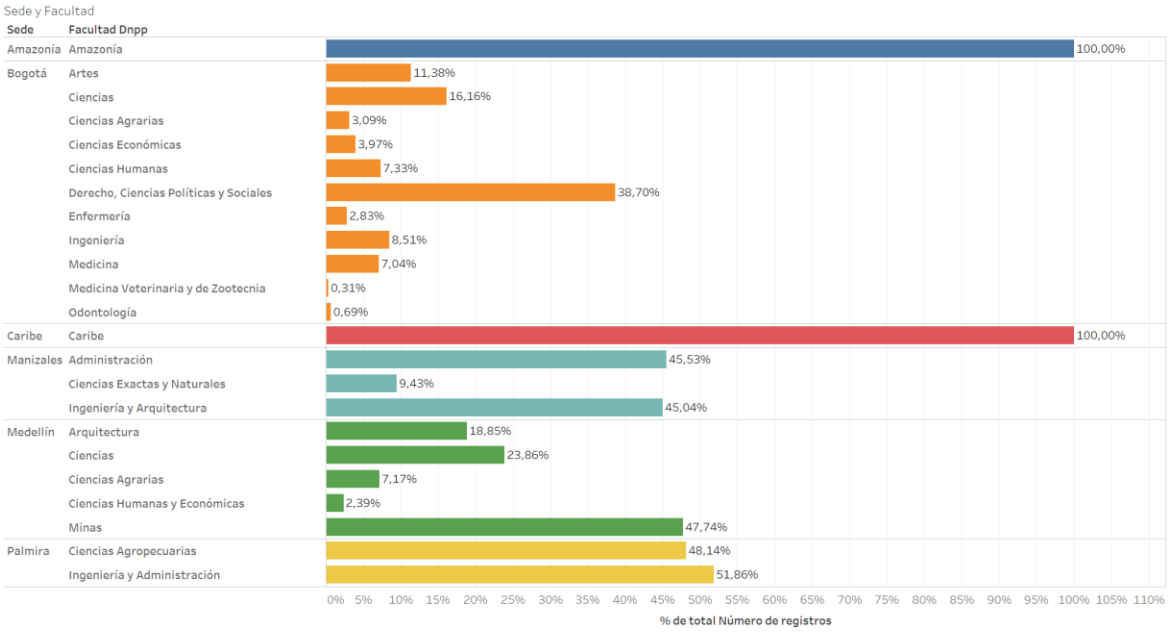
Una vez se tienen los datos de los estudiantes que se graduaron, se hizo un cruce con los datos que esas personas tenían en la base de matriculados, y en la base de admitidos para poder lograr tener la mayor cantidad de variables posibles para cada uno de los registros. Al obtener estos datos, se procedió a hacer un análisis descriptivo, que se realizó con ayuda del software de análisis e inteligencia de negocios *Tableau*[4]. Vale la pena anotar que, aunque este software no es de código abierto, permite a los estudiantes y profesores (con previo registro) obtener una licencia estudiantil durante 1 año.

A pesar de que la Universidad Nacional tiene 8 sedes, solo tiene programas de posgrado en 6 de ellas. En la Figura 2-3 se puede observar cómo se divide el número de graduados por sede y facultad, y se puede observar como en la sede Palmira el porcentaje de graduados es casi igual para las 2 facultades. Un caso similar pasa en Manizales, donde la Facultad de Ingeniería y Arquitectura, y la de Administración tienen el 91% de los graduados. Por lo cual, dividiendo este 91% en porcentajes casi iguales, estos son cercanos al 45%. En Medellín y Bogotá ocurren casos distintos: para ambas sedes se observa como 2 facultades tienen un gran porcentaje de los graduados: la facultad de Minas en Medellín con un 47.7%, y la facultad de Derecho en Bogotá con un 38.7 %. Esto se puede explicar con base en el número de programas que tienen esas facultades, ya que

[4] Análisis e inteligencia de negocios – *Tableau Software,* Tomado de: https://www.tableau.com/es-es

como se puede observar en la Figura 2-1, los programas que más especializaciones tienen son Minas en Medellín y Derecho en Bogotá.

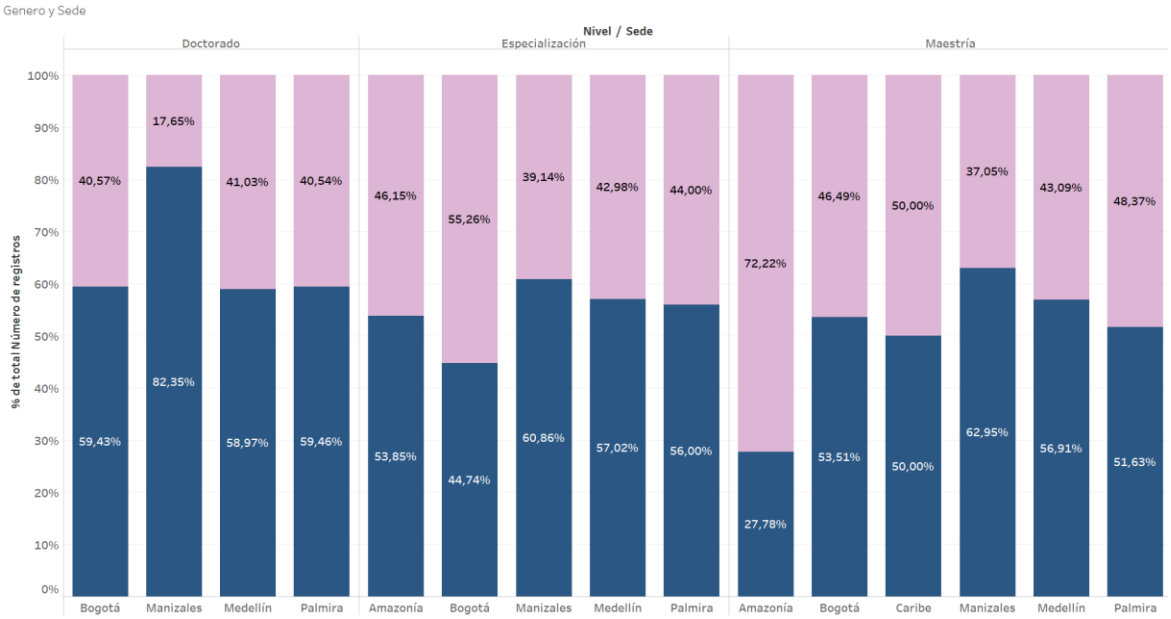
Figura 2-3: Distribución de graduados por Sede y Facultad



Otro variable sobre la que se puede hacer un análisis es el género de las personas graduadas, para esto como se puede ver en la Figura 2-4, se dividió el número de graduados por sede y nivel, para poder hacer el análisis, ya que si, no se hace un filtro inicial, el número de hombres y mujeres graduados es muy similar, pero al dividirlo por sede y nivel, se pueden observar que para los 3 niveles, el número de hombres graduados en Manizales en promedio es de 68.72%, lo que, en principio, hace pensar que en Manizales se debería implementar una estrategia para hacer que las mujeres puedan lograr graduarse en mayor número. En Amazonía, se presenta la tendencia contraria para las maestrías, el 72.22 % de las personas que se graduaron son mujeres, cabe aclarar que en la sede Amazonía solo hay una maestría, y que con este dato lo que se debería hacer es una mayor promoción del programa o tratar de crear asignaturas que puedan atraer el interés para que más hombres ingresen al programa. Cabe anotar que estas estrategias se plantean como posibles recomendaciones a partir de esta investigación, para buscar que haya equidad entre hombres y

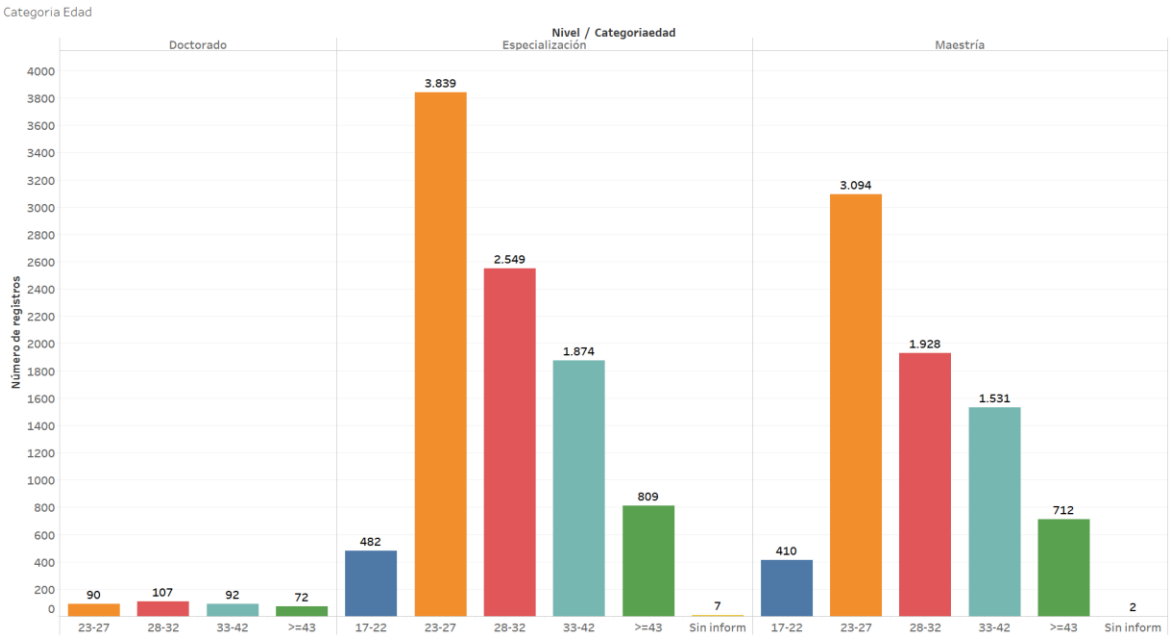
mujeres. En el resto de sedes y niveles, se presentan unas cifras relativamente iguales entre hombres y mujeres en lo relacionado a la tasa de graduación.

Figura 2-4: Número de graduados por sexo



Otra clasificación que se pudo hacer fue con respecto a la edad; en la Figura 2-5 se puede observar cómo se hizo una agrupación de la variable edad para tratar ciertas categorías y poder obtener una mejor información sobre los graduados. En esta gráfica se puede ver que la categoría de mayor cantidad de personas para maestría y especializaciones es la que comprende estudiantes con edades entre 23 y 27. Vale resaltar que la mayoría de estudiantes comienzan sus estudios de posgrado, una vez terminan su pregrado, o esperan uno o dos años para hacerlo. Para Doctorado, pasa una situación similar, la mayoría de personas comienzan el Doctorado cuando están en el rango comprendido entre 28 y 32 años, usualmente después de terminar la maestría.

Figura 2-5: Clasificación de la edad de los graduados por nivel académico de posgrado

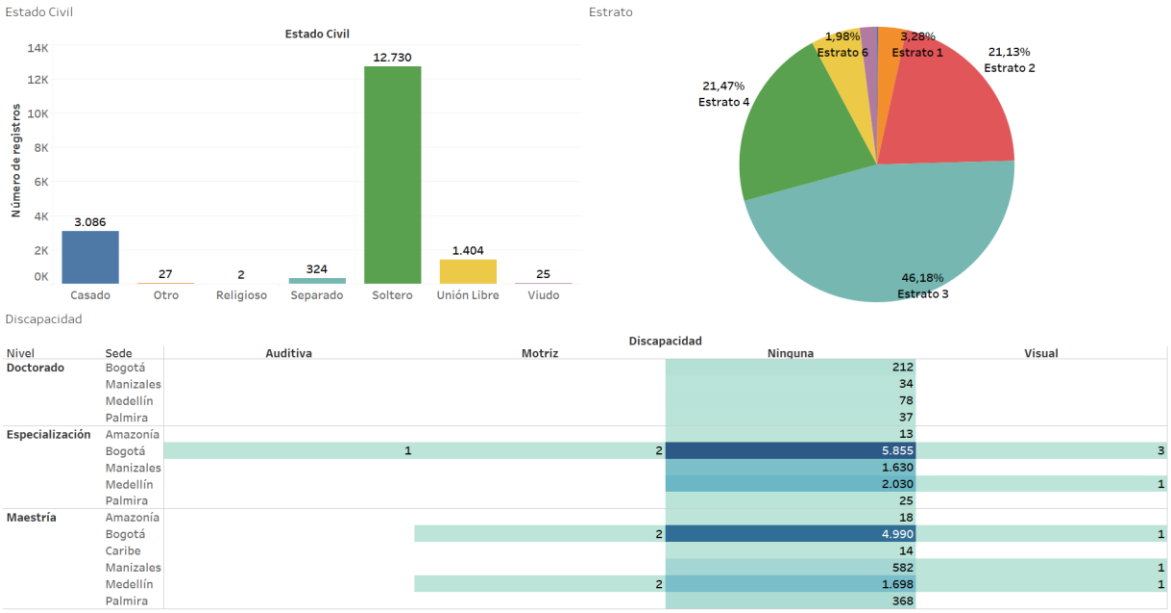


En cuanto a variables sociales y personales de cada graduado, en la Figura 2-6 se pueden observar el compendio de 3 de ellas: estrato, estado civil, y discapacidad. En cuanto al estrato, se puede observar que los que ingresan a estudiar a la Universidad Nacional son personas que están entre los estratos 2, 3 y 4. Un 88.78% pertenecen al estrato 3, el que más graduados tiene: un 46.18%. Vale anotar que las cifras que actualmente se manejan[5] son coherentes con el 62.9% de habitantes de Bogotá que viven en los estratos 3 y 4, y el 34.8% que lo hacen en estratos 1 y 2. En cuanto al estado civil, cuando inician un posgrado, se puede observar que hay un 26% que reporta tener una vida en pareja, particularmente, casados y en unión libre; pero, la mayoría (72.4%) de los que ingresan a un programa de posgrado son personas solteras. Por último, en la variable discapacidad, se puede observar como muy pocas personas de las que se han graduado informa tener alguna discapacidad, y entre esas personas, la mayoría (9 personas de 14) se graduaron en Bogotá en los programas de maestría y especialización. A nivel general, no se tiene información de ninguna

[5] La mitad de los habitantes de Bogotá son de clase media, El Espectador, 2017, Tomado de: https://www.elespectador.com/noticias/bogota/la-mitad-de-los-habitantes-de-bogota-son-de-clase-media-articulo-707875

persona con alguna discapacidad (auditiva, motriz o visual) que se haya graduado de algún programa de Doctorado.

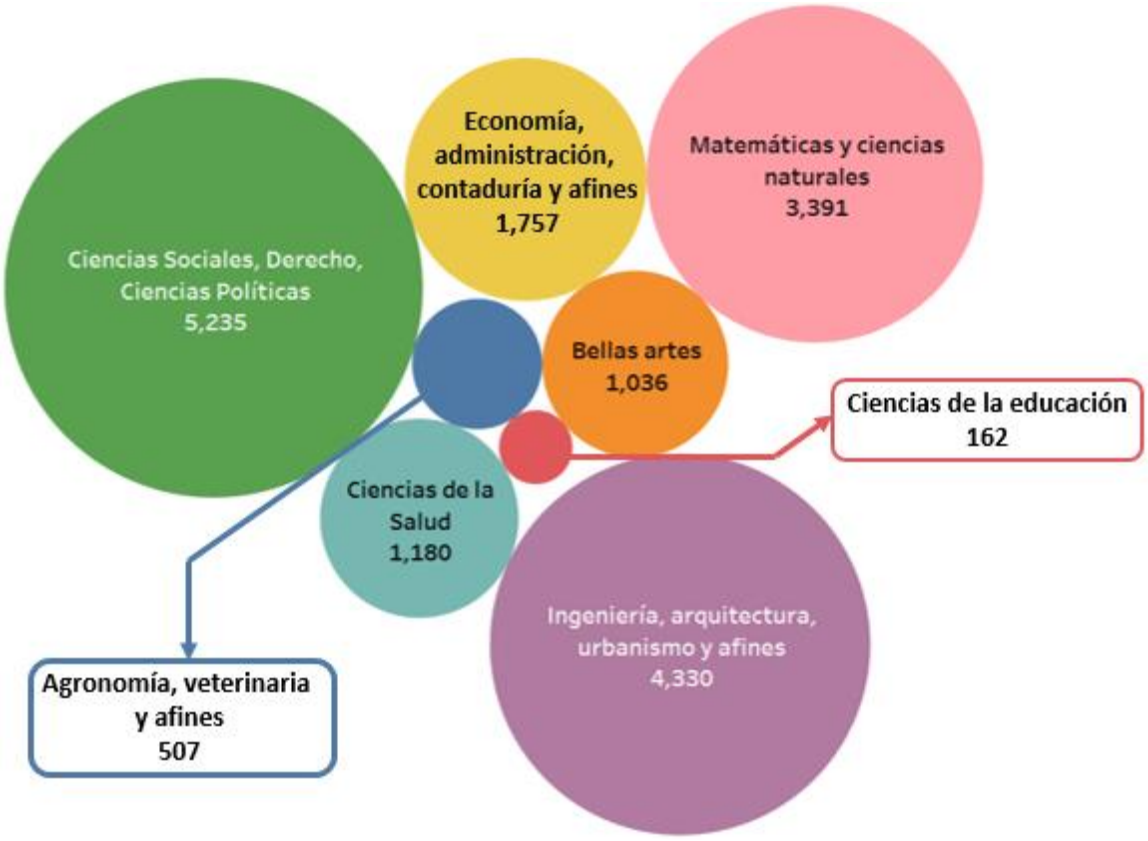
Figura 2-6: Atributos Estado Civil, Estrato y Discapacidad para los graduados



Nivel	Sede	Auditiva	Motriz	Ninguna	Visual
Doctorado	Bogotá			212	
	Manizales			34	
	Medellín			78	
	Palmira			37	
Especialización	Amazonía			13	
	Bogotá	1	2	5.855	3
	Manizales			1.630	
	Medellín			2.030	1
	Palmira			25	
Maestría	Amazonía			18	
	Bogotá		2	4.990	1
	Caribe			14	
	Manizales			582	1
	Medellín		2	1.698	1
	Palmira			368	

Ahora teniendo en cuenta el programa del que se graduaron se puede ver como los programas pertenecientes al área de conocimiento de Ciencias Sociales, Derecho y Ciencias Humanas tienen 5235 graduados, presentando una gran diferencia con la siguiente área de conocimiento con 4330; esto se puede explicar por el hecho de que la Facultad de Derecho de la sede Bogotá, tiene 12 especializaciones, por lo cual, esta facultad aporta el mayor número de graduados para esta área de conocimiento. En cuanto a ingeniería, arquitectura, urbanismo y afines (la segunda con mayor cantidad de graduados), los programas de las facultades de Ingeniería, Artes y Arquitectura que tienen programas en esta área de conocimiento, también contribuyen con un buen número de graduados. Una diferencia observada es que, a pesar de que tienen un número mayor de especializaciones, no tienen tantos graduados en estas como si los tiene la facultad de Derecho.

Figura 2-7: Área de conocimiento del programa del que se graduó cada estudiante



En la Universidad Nacional de Colombia existen 2 tipos de planes de estudio, Investigación y Profundización, clasificación disponible para maestrías; los Doctorados son programas de investigación y las especializaciones, aunque se podrían considerar como programas de profundización, por las normas de la Universidad estos programas no tienen esta clasificación. En la Figura 2-8 se puede observar entonces como hay un número similar en la cantidad de graduados de maestría en las dos modalidades (profundización e investigación), con un leve sesgo hacia el lado de las maestrías de Investigación con el 53.7% de los graduados. Esto se podría explicar porque existen muchas más maestrías que solo tienen el plan de Investigación, en comparación con los programas que tienen o ambos tipos de plan de estudios, o los que solo tienen profundización. Por tal motivo, la cantidad de graduados es mayor para las maestrías en Investigación.

Por otro lado, en la Figura 2-9 se puede observar cómo se distribuye la población de los graduados según su ciudad de nacimiento. Se puede observar que las ciudades principales ponen la mayor cantidad de persona graduadas, siendo Bogotá la que más graduados tiene con el 32.2%, seguida por Medellín con el 11.8%; las ciudades no capitales que más graduados tienen son Palmira y Duitama, esto se puede explicar por su cercanía a las sedes de Palmira y Bogotá, respectivamente.

Figura 2-8: Distribución de los graduados por tipo de plan de estudios y nivel

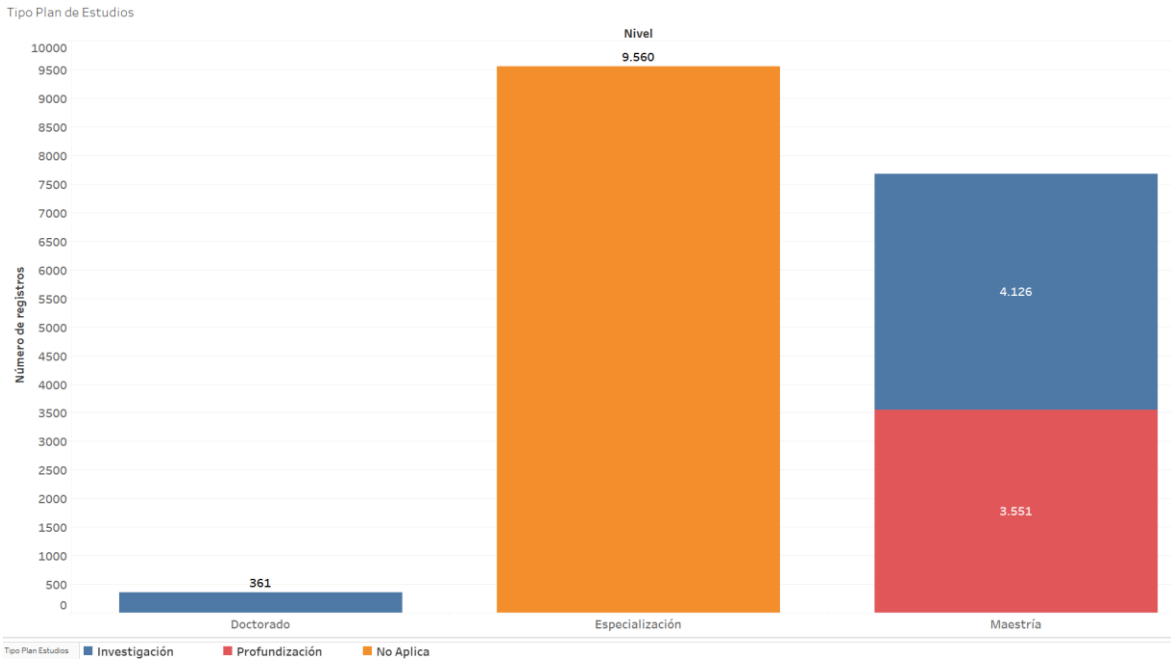
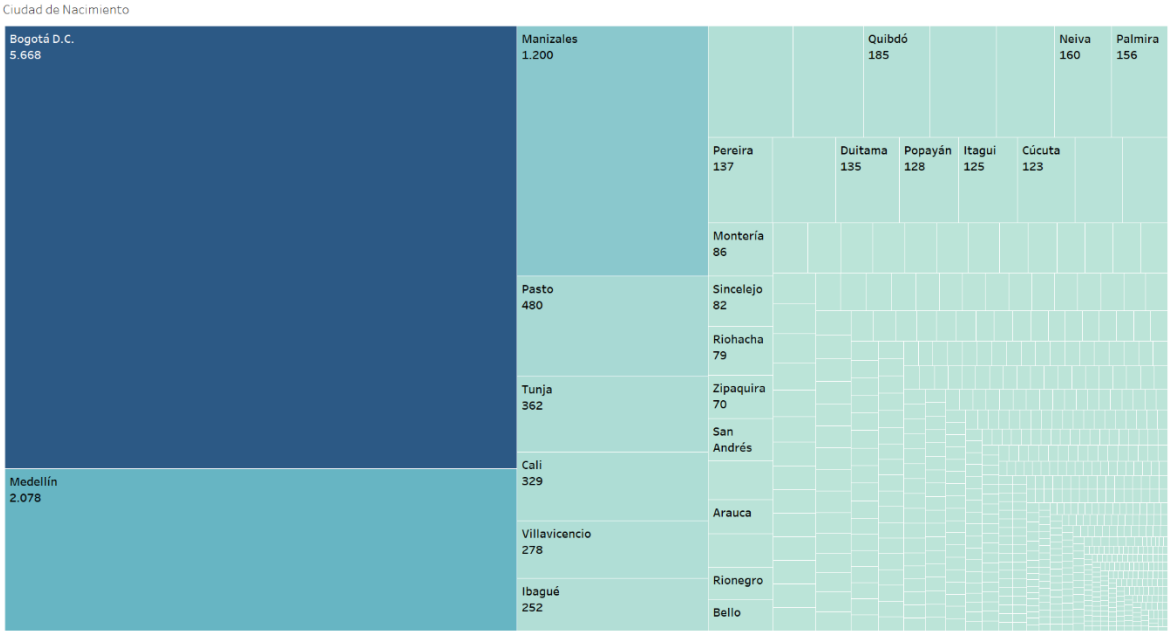


Figura 2-9: Mapa del árbol[6] de la variable Ciudad de Nacimiento



En las figuras 2-10, 2-11 y 2-12 se presenta la información del número de graduados, divididos por sede, facultad y nivel, también mostrando la división con respecto a los convenios. En la figura 2-10 se muestra como ninguna otra facultad tiene tantos graduados como los programas de especialización de la Facultad de Derecho de la sede Bogotá, esta facultad tiene una cifra de 3906 graduados, se puede observar que más de la mitad de los graduados lo han hecho por medio de algún convenio con otras instituciones (color azul). Otros programas con convenios a destacar son la especialización de la sede Amazonía (ver Figura 2-11), y las maestrías de la Facultad de Administración de la sede Manizales (Ver Figura-12), la diferencia es que los convenios de la Facultad de Derecho aportan el 11.59% de los graduados con respecto al total de la población de estudio. También se observa que la Facultad de Ingeniería, que es una de las que más programas tiene, apenas aporta aproximadamente 1000 graduados en los últimos 8 años. Para las demás sedes, se observa que en algunas facultades, los convenios tienen un pequeño porcentaje de los graduados que oscila entre el 5% y el 20% dependiendo de la facultad, pero no tienen un comportamiento tan particular como el descrito anteriormente.

[6] Como crear un diagrama de árbol. Disponible en: https://onlinehelp.tableau.com/current/pro/desktop/es-es/buildexamples_treemap.html

Figura 2-10: Tipo de convenio dividido por sede y nivel académico de posgrado (Sede Bogotá)

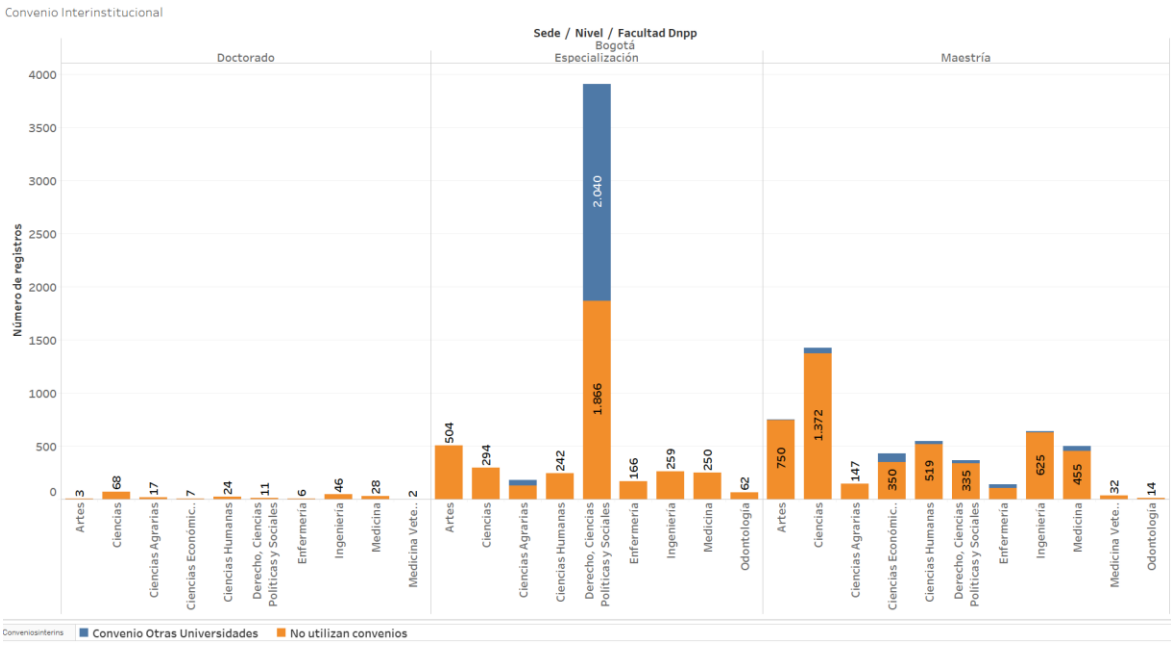


Figura 2-11: Tipo de convenio dividido por sede y nivel académico de posgrado (Sede Amazonía, Sede Caribe y Sede Palmira)

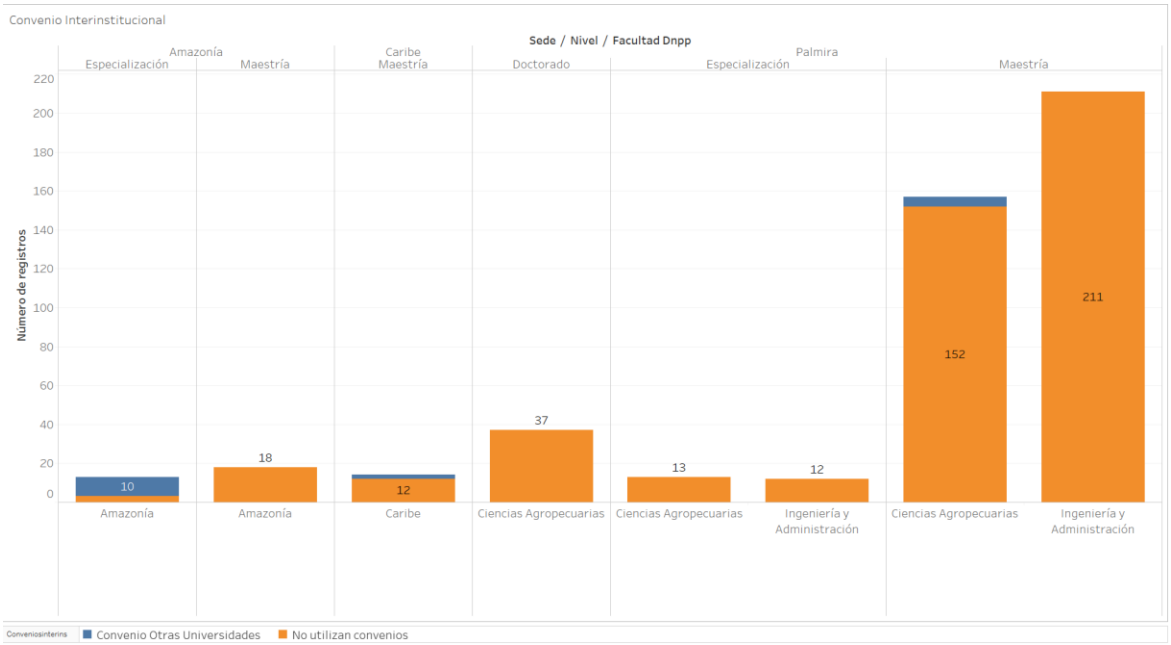
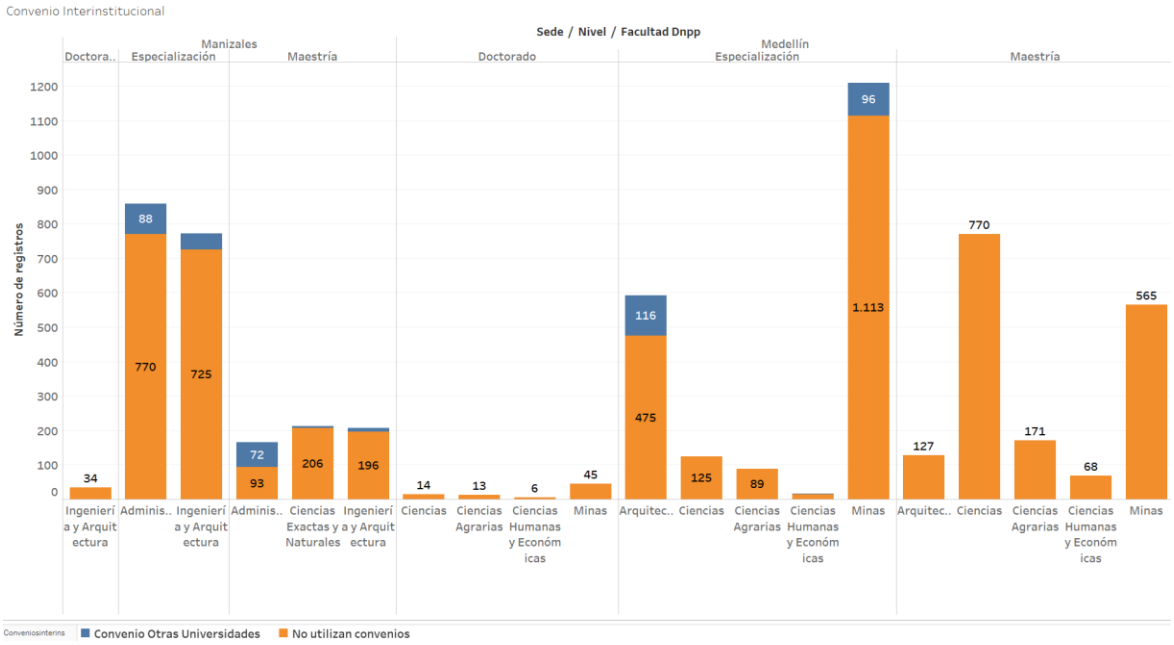
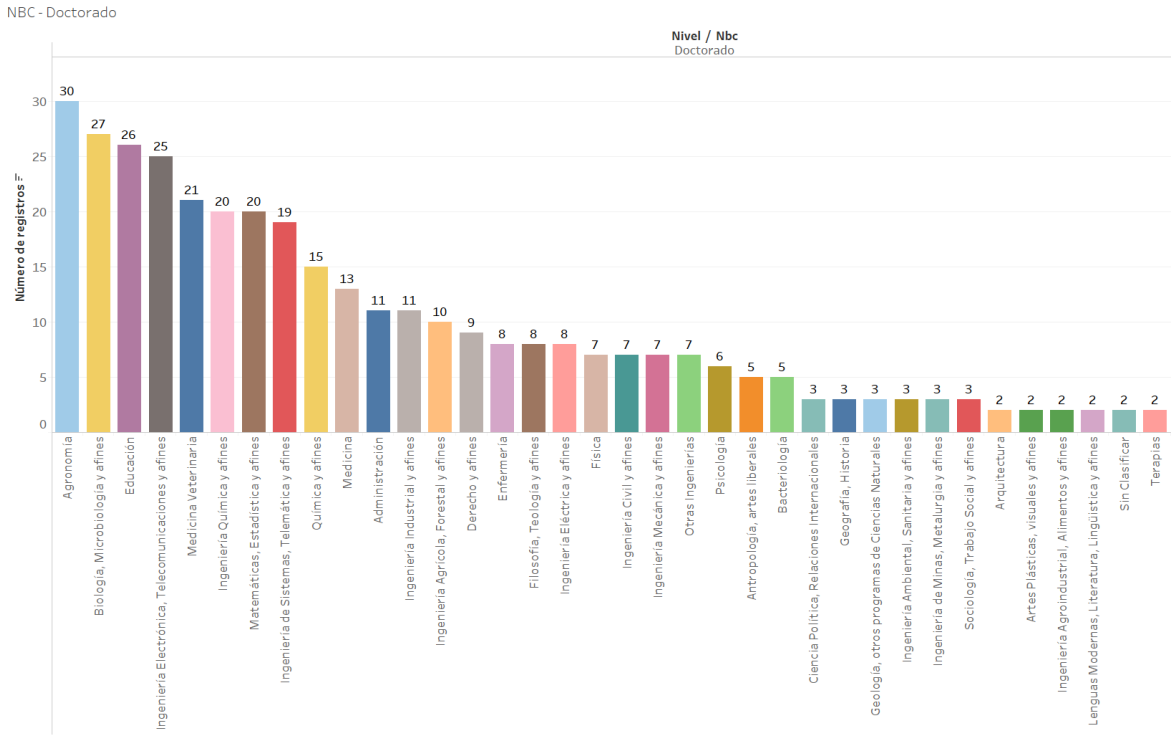


Figura 2-12: Tipo de convenio dividido por sede y nivel académico de posgrado (Sede Manizales y Sede Medellín)



Finalmente, para poder realizar una agrupación con base en la variable del título de pregrado obtenido por cada graduado, se acudió a utilizar el Núcleo Básico del Conocimiento (NBC), que es una división o clasificación utilizada por el Ministerio de Educación de Colombia, en el cual se dividen las áreas de conocimiento en campos, disciplinas o profesiones más específicas ya que describen de mejor manera los programas que lo componen. En la Figura 2-13 se observa una cierta homogeneidad en la cantidad de graduados de cada campo y, que aunque entre los de mayor valor se pueden destacar Agronomía, Biología, Microbiología y afines, y Educación, en general, todos los campos de estudio tienen valores muy cercanos.

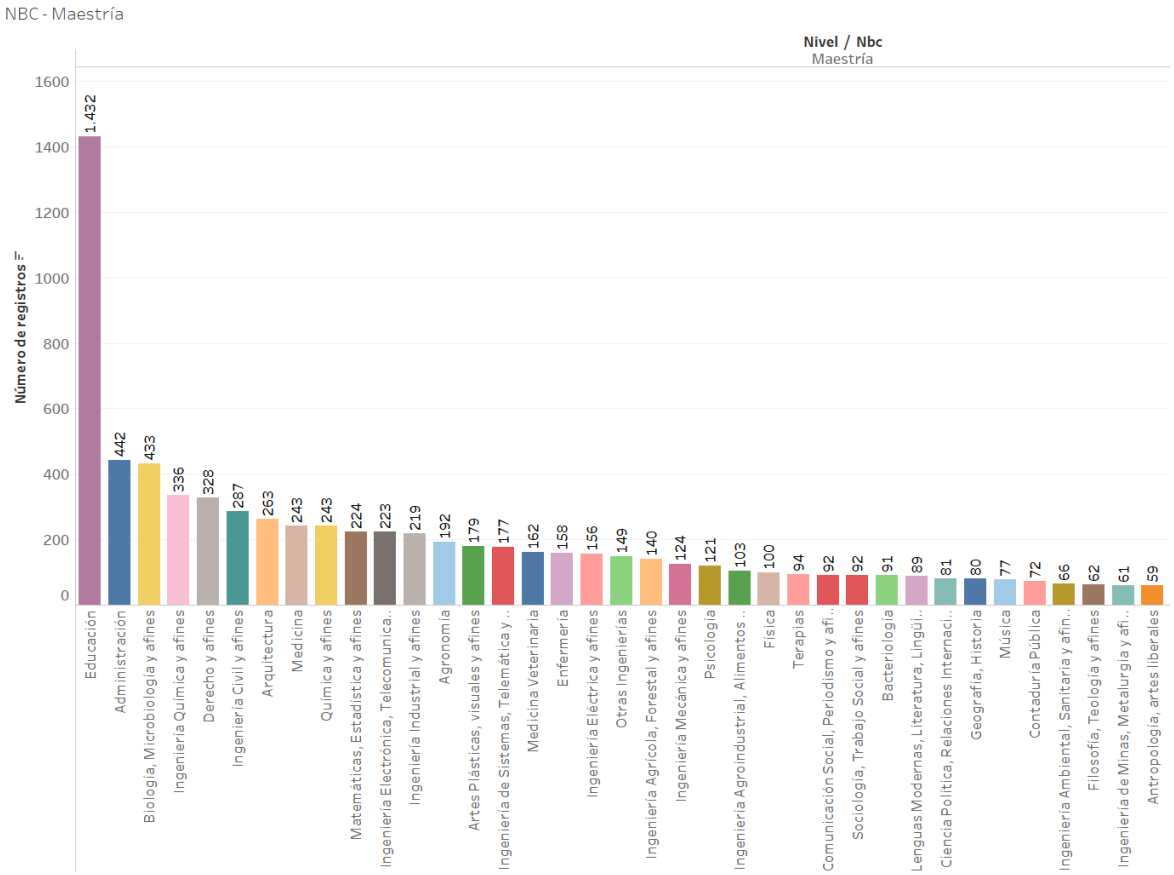
Figura 2-13: Distribución de los graduados por Núcleo Básico del Conocimiento (Nivel de Doctorado)



En la Figura 2-14 se observa la distribución de los NBC para los programas de maestría. En esta gráfica se puede observar un campo que sobresale, el campo de la Educación, con 1432 graduados. Esta cifra se podría explicar, en un programa [7] que tiene el Ministerio de Educación en el cual le da posibilidades y facilidades a los docentes de Bogotá para estudiar alguna maestría. Y aunque los graduados de este campo se dividen en varios programas, el que más cantidad de graduados tiene es la Maestría en Enseñanza de las Ciencias Exactas y Naturales presente en Bogotá, Palmira, Manizales y Medellín con 941 graduados.

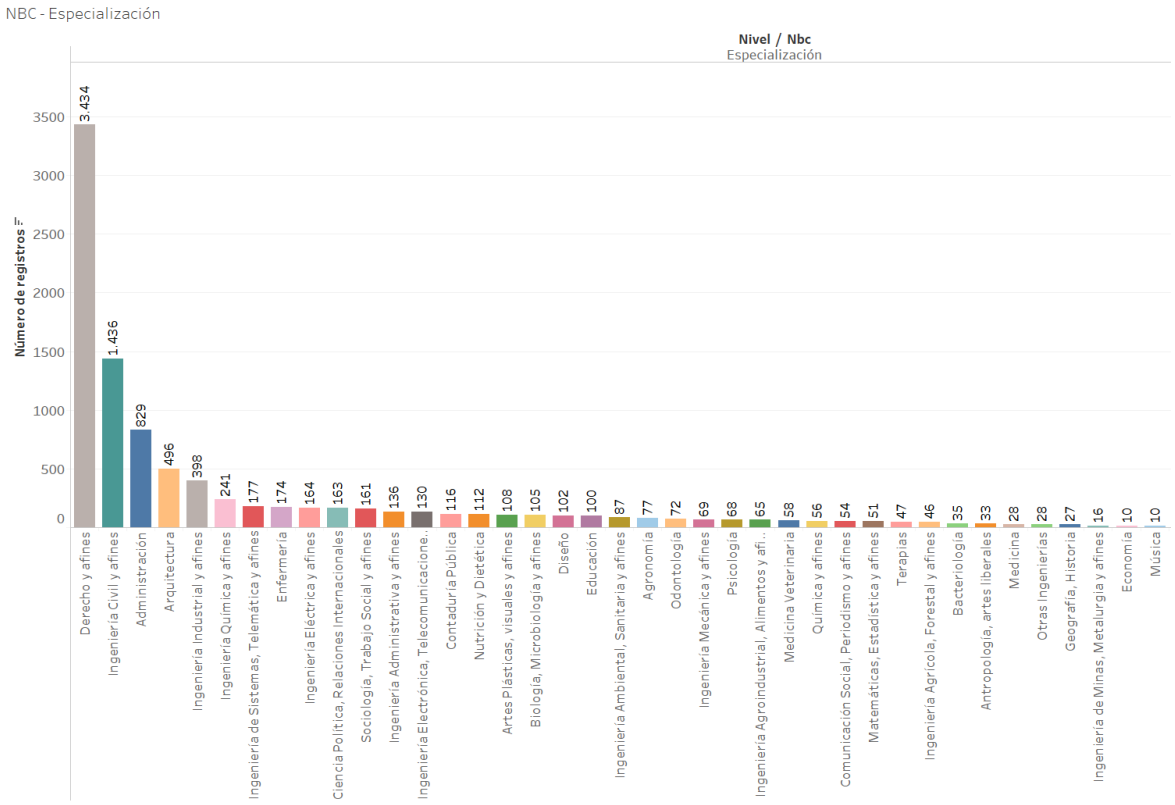
[7] Fondo para la formación posgradual de docentes. Tomado de: https://www.icetex.gov.co/dnnpro5/Default.aspx?tabid=1188

Figura 2-14: Distribución de los graduados por Núcleo Básico del Conocimiento (Nivel de Maestría)



Como ya se ha visto a lo largo de este análisis descriptivo, el NBC de Derecho y afines es el que más graduados tiene a nivel de especializaciones con 3434. Esta cifra no solo tiene que ver con los convenios, también se puede ver que los graduados de Derecho de las diferentes universidades, buscan estudiar una especialización, tal vez con el ánimo de tener una ventaja en el mercado laboral. En el segundo puesto está el NBC de Ingeniería Civil y afines; esto se debe también a que las facultades de Ingeniería y Arquitectura en Manizales, Minas y Arquitectura en Medellín e Ingeniería en Bogotá ofertan varios programas que son de interés para los ingenieros civiles, y que al igual que los graduados de Derecho, y, en general, muchos de los graduados de pregrado ven las especializaciones como una forma de tener una ventaja en el mundo laboral, y de poder profundizar sus conocimiento en algún campo de su área que sea de interés.

Figura 2-15: Distribución de los graduados por Núcleo Básico del Conocimiento (Nivel de Especialización)



Finalmente, dentro de otras variables que se tuvieron en cuenta para el modelo, se encontró por ejemplo, que la mayoría de los que se graduaron ingresaron al programa por medio de admisión regular, o que el 41.29% de los graduados hicieron su pregrado en alguna de las sedes de la Universidad Nacional de Colombia, lo que podría significar que muchos de los que se gradúan de pregrado de la Universidad se presentan a un posgrado en la misma Universidad, o también se puede pensar que estas personas ya conocen cómo es estudiar en la Universidad Nacional, por lo cual, se les podría hacer más conveniente poderse graduar de un programa, en contraste con personas que provienen de otra Universidad y podrían presentar dificultades para graduarse por, tal vez, requerir mayor acompañamiento por parte del programa.

.

3. Análisis de asociación y generación del modelo

En este capítulo se describe el proceso que se llevó a cabo para generar un modelo que permitiera obtener las relaciones de asociación de los datos de las personas graduadas.

3.1 Preprocesamiento

Una de las fases importantes que tiene el proceso de KDD es precisamente el pre-procesamiento que se debe llevar a cabo con los datos originales para que se puedan aplicar posteriormente los algoritmos necesarios. Para este caso los datos se obtuvieron básicamente de varias fuentes de la Universidad, la proporcionada por la Dirección Nacional de Admisiones (DNA) en donde se tiene toda la información de las personas inscritas y admitidas a la Universidad, que también tiene datos sociales como el estrato, estado civil y discapacidad. La segunda fuente de información que se tuvo en cuenta fue el Sistema de Información Académica (SIA), en el cual se registra toda la trazabilidad de los estudiantes a lo largo de sus estudios; aparte de la información de matrículas, también se pueden observar datos de los bloqueos académicos y no académicos, que pueden llevar a los estudiantes a desvinculaciones temporales o definitivas. Una de las fuentes principales es la Secretaría Académica de cada facultad, ya que esta dependencia tiene consolidada la información de las personas que logran graduarse de un programa de la Universidad Nacional, por lo cual, se tomó la información proporcionada por esta dependencia como base para esta investigación.

Por otro lado, la Dirección Académica en la Sede Bogotá es la que administra las Becas de Asistente Docente, que son becas que se les proporcionan a estudiantes de maestría para cubrir la mayor parte de su matrícula, a cambio de que los estudiantes se apoyar el desarrollo de uno o dos cursos ofrecidos a estudiantes de pregrado. Otra beca que se tuvo en cuenta fue la que se maneja en la Dirección Nacional de Programas de Posgrado, la Beca otorgada por Colciencias, en la cual de

acuerdo con los requisitos establecidos por Colciencias, se cubre la totalidad de la matrícula para estudiantes de Doctorado, lo que les da muchas facilidades para dedicarse de tiempo completo a adelantar sus estudios.

Entonces una de las primeras tareas que se debió realizar fue hacer lo que en bases de datos relacionales se conoce como un JOIN, es decir, unir varias tablas, cada una de ellas determinada por un identificador, para poder tener una sola base de datos con la mayor cantidad de información que se pudiera obtener sobre cada graduado. Toda la información se registró en una base de datos MySQL para facilitar los procedimientos que se debían realizar a continuación.

Una vez se tenía la base de datos con todas las variables de las diferentes fuentes de información, se procedió a eliminar las que no le aportarían valor al modelo, por ejemplo, los nombres y apellidos de la persona, algunos identificadores de los distintos campos o variables repetidas que ya se tenían de otras bases de datos, entre otras.

Así como se hizo reducción de columnas, también se procedió a realizar reducción de filas, para esto en primer lugar se buscaron los duplicados y se dejaron los que tuvieran la información más completa. También, haciendo un análisis de cada variable se detectaron los registros que no tuvieran la información completa, y si, además, en varias variables de estos registros se presentaba la misma situación, se procedió a eliminar dichos registros.

Otro paso que se realizó para la limpieza de datos fue hacer la corrección sobre los datos de cada variable, es decir, corregir la ortografía, convertir las mayúsculas a minúsculas, editar el formato de la fecha de nacimiento, ya que se tenían registros en los que se presentaba la fecha de nacimiento de la forma YY/MM/DD (año/mes/día), pero otros tenían DD/MM/YY (día/mes/año), entonces lo que se hizo fue dejar todas las fechas de nacimiento en el mismo formato, para realizar el cálculo de la variable edad.

Después de esta limpieza preliminar se pudo obtener un conjunto de variables académicas y no académicas para caracterizar al programa o al estudiante. En la Tabla 3-1 se pueden observar las

variables que identifican a cada programa, en la Tabla 3-2 las variables no académicas que identifican a cada persona, y en la Tabla 3-3 las variables académicas.

Tabla 3-1: Variables que caracterizan al programa

Nombre Variable	Tipo de Datos	Variación de Datos
Sede	Texto	6 Sedes
Facultad	Texto	21 Facultades
Nivel	Texto	3 Niveles
Nombre Programa	Texto	279 Programas
Área de Conocimiento	Texto	8 Áreas de Conocimiento
Estado de Acreditación	Texto	6 Estados de Acreditación

Tabla 3-2: Variables no académicas que describen a cada graduado

Nombre Variable	Tipo de Datos	Variación de Datos
Genero	Texto	{Masculino, Femenino}
Ciudad Nacimiento *	Texto	760 Ciudades
Universidad de Pregrado	Texto	Datos variados
Título Pregrado	Texto	Datos Variados
Edad Calculada	Numérico	Datos variados
Categoría de Edad	Texto	5 Categorías de edad calculadas por la distribución de los datos
Discapacidad	Texto	4 Estados
Estado Civil	Texto	7 Estados Civiles
Estrato	Texto	7 Estratos

Tabla 3-3: Variables académicas que caracterizan a cada graduado

Nombre Variable	Tipo de Datos	Variación de Datos
Tipo de Solicitud de Admisión	Texto	4 tipos de admisión
Convenio	Texto	Si o No
Institución de Convenio	Texto	35 Estados posibles
Tipo Plan de Estudios	Texto	Investigación o Profundización
Beca	Texto	Si o No
Tipo Beca	Texto	2 estados posibles (Colciencias o Asistente Docente)
No. Periodos para completar ciclo de estudios	Numérico	Variado desde 2 hasta 17 semestres
No. Periodos para graduarse	Numérico	Variado desde 3 hasta 17 semestres
No. Matriculas	Numérico	Variado desde 1 hasta 9
P.A.P.A	Numérico	Variado

Una vez se tienen estas variables, se realizó una revisión de algunas de ellas para poder hacer agrupación de aquellas que presentaron valores muy diversos y que se visualizaba que tenían alguna posibilidad de transformarse en una variable, con mucho menos cantidad de datos, por ejemplo, las variables universidad de pregrado y título de pregrado. Con la variable universidad de pregrado, lo que se hizo fue una limpieza de todos los datos tratando de disminuir los valores posibles para este campo. Luego se revisó el nombre de cada universidad y se estandarizó de manera que no se consideraran como diferentes nombres de universidades con valores correspondientes a la misma universidad y que fueran diferentes, por ejemplo, por una tilde, por un símbolo o por una abreviación. Realizando este procedimiento se logró reducir el número de valores posibles de 440 a 180. Para la variable título de pregrado se llevaron a cabo dos procedimientos, el primero similar al que se hizo con universidad de pregrado, en donde se logró reducir de 1100 valores posibles a tan solo 315. El segundo procedimiento hace referencia a la

creación de una nueva variable para tratar de agrupar esos 315 títulos; se encontró entonces que el Ministerio de Educación hace una agrupación de estos programas, llamada Núcleo Básico del Conocimiento (NBC). Por consiguiente, con las clasificaciones del Ministerio de Educación se empezó a clasificar cada uno de los títulos en su NBC correspondiente, obteniendo al final que se redujeron 315 títulos a 53 NBC.

3.2 Generación de clasificación de registros

Los programas de posgrado de la Universidad Nacional de Colombia tienen la particularidad de tener duraciones diferentes en momentos distintos, aunque cada modificación al respecto debe ser aprobado por los cuerpos colegiados respectivos. Esta situación provocó un inconveniente para la generación de una clasificación, porque un estudiante que se graduó hace 5 años en un programa X, puede que el programa haya estimado la duración en 3 años, pero un estudiante que se graduó hace un año en el mismo programa X, puede que lo haya hecho cuando la duración del programa estaba ahora reglamentada en 4 años.

Entonces el procedimiento para resolver este inconveniente fue hacer una trazabilidad de la duración de cada programa, desde 2008 hasta el segundo semestre de 2016, revisando los Acuerdos del Consejo Académico por el cual se modificaba la duración de los diferentes programas. De esta manera se obtuvo un listado de la duración de cada uno de los programas a través de los años. Una vez se tenía ese dato, se asoció la duración de cada programa al registro de cada estudiante, teniendo en cuenta el semestre en el que iniciaron a estudiar, y de ese modo teniendo la duración reglamentada dR y el tiempo en la que cada graduado completó el ciclo de estudios $tCCE$, ahora si es posible asignar una etiqueta a cada uno de los registros.

Como desde un principio la pregunta de investigación planteada era identificar las causas para que un graduado no obtuviera su título en el tiempo reglamentado por cada programa. También, siguiendo la línea de la normativa de desvinculación temporal que habla de exceder el tiempo máximo de permanencia (siempre es del doble de la duración reglamentada por cada programa),

entonces, en conjunto con la Dirección Nacional de Programas de Posgrado, se establecieron 4 categorías para clasificar los registros:

- **Menor a la duración del programa:** muchas personas que van terminando sus estudios de pregrado ya tienen en mente iniciar un posgrado, generalmente una Maestría, entonces lo que hacen esos estudiantes es empezar a cursar asignaturas del posgrado como asignaturas de libre elección en su pregrado; con esto logran que una vez sean admitidos al posgrado puedan homologar estas materias y ya solo quedarían pendientes de cumplir con las actividades académicas, es por este motivo que esta categoría tiene sentido. Para que un registro esté en esta categoría debe cumplir que:

$$Tiempo\ para\ Completar\ Ciclo\ de\ Estudios < Duración\ del\ programa$$

- **Igual a la duración del programa:** es el ideal, cada uno de los programas hace un análisis de sus planes de estudio y los requisitos establecidos para que una persona se gradúe, y así establecer una duración acorde a ello. El filtro que se hace a los registros para colocar esta etiqueta es:

$$Tiempo\ para\ Completar\ Ciclo\ de\ Estudios = Duración\ del\ programa$$

- **Entre la duración del programa y el doble de la duración del programa:** los programas de posgrados ofrecen una facilidad para los estudiantes, la cual consta de aplazar sus estudios en máximo dos oportunidades, cada una con 1 semestre de duración, durante el tiempo que se encuentre estudiando. Muchos estudiantes toman esta opción ya sea por alguna razón personal, económica, o para tomar ese tiempo para adelantar sus trabajos de grado, sin tener otras obligaciones académicas de por medio. La condición para pertenecer a éste grupo es:

$$dP < tCCE \leq 2dP$$

- **Mayor al doble de la duración del programa:** las personas que ya han cumplido el doble de la duración del programa y no han logrado graduarse, automáticamente son desvinculadas de la Universidad. Sin embargo, tienen la oportunidad de culminar sus estudios solicitando un reingreso al Consejo Superior Universitario, el cual, si es aprobado, solo tiene duración de un semestre y el estudiante lo debe utilizar para terminar y entregar a la facultad respectiva su tesis o trabajo de grado terminado, para que así pueda conseguir

graduarse de un programa de posgrado. Lo que hacen parte de este grupo son lo que cumplen con:

$$Tiempo\ para\ Completar\ Ciclo\ de\ Estudios > 2 * Duración\ del\ Programa$$

Una vez cada uno de los registros tenía esta clasificación, se pudieron identificar algunas problemáticas y obtener algunas conclusiones interesantes. En la Figura 3-1 se observa como, a nivel general, solo el 38.41% de los que se graduaron lo hicieron en el tiempo reglamentado por cada programa. Por el contrario, la mayoría de los estudiantes lo ha hecho entre la duración del programa y el doble del mismo. Vale la pena anotar que, en la Figura 3-2, las especializaciones fueron las que tenían la mayoría de los registros que se clasificaban como "Igual a la duración del programa". Pero, para las maestrías (ver Figura 3-3) y para los doctorados (ver Figura 3-4), la mayoría de los registros están "Entre la duración del programa y el doble", con 68.74% y 79.50% respectivamente. El segundo lugar es para "Mayor al doble de la duración del programa", con un 19.14% para maestrías y un 13.3% para doctorados. Entonces, observando estos datos preliminares se confirma la problemática planteada, y la necesidad de que los programas empiecen a desarrollar estrategias académicas y de bienestar para mejorar estas cifras, sobre todo para los programas de Doctorado y Maestría, en ese orden.

Figura 3-1: Distribución de las etiquetas de clasificación en el conjunto de datos

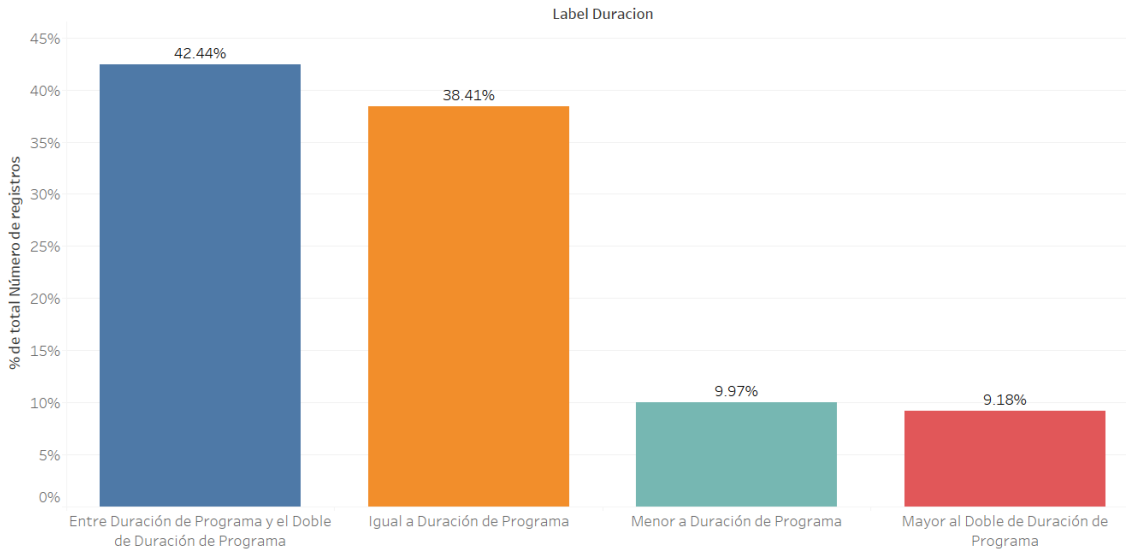


Figura 3-2: Distribución de las etiquetas de clasificación a nivel de Especialización

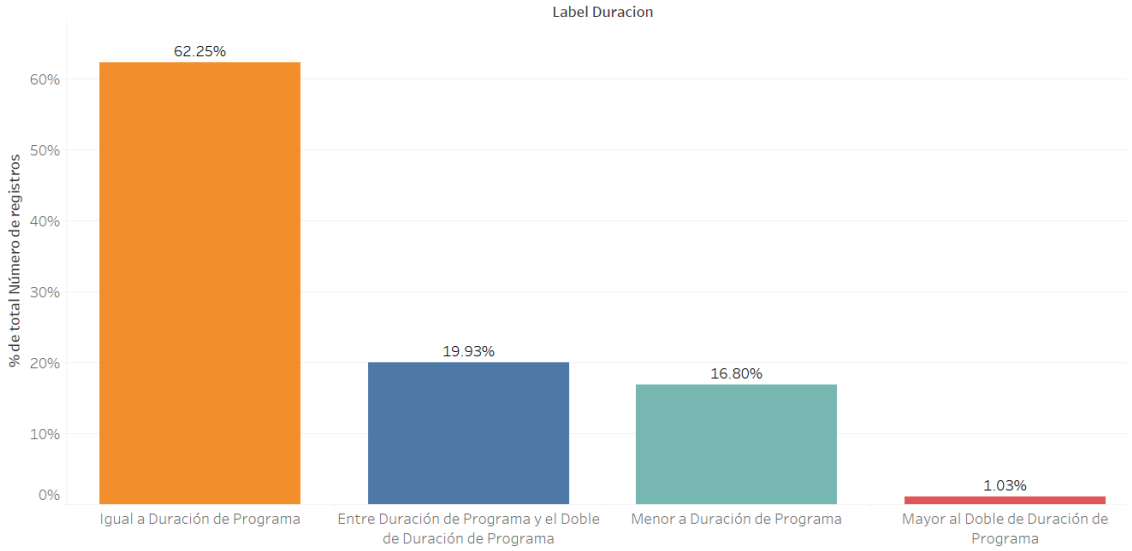


Figura 3-3: Distribución de las etiquetas de clasificación a nivel de Maestría

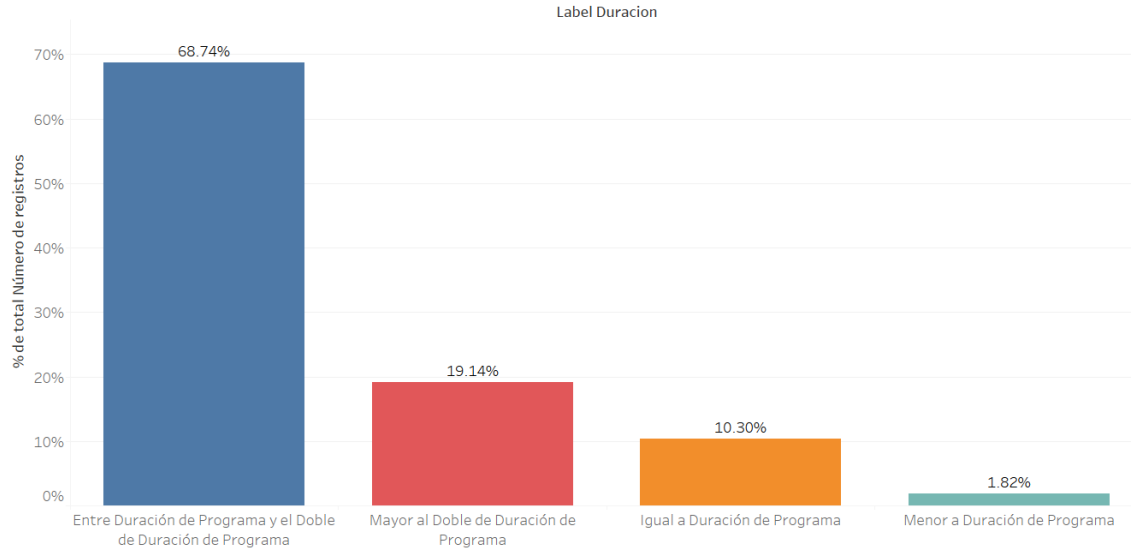
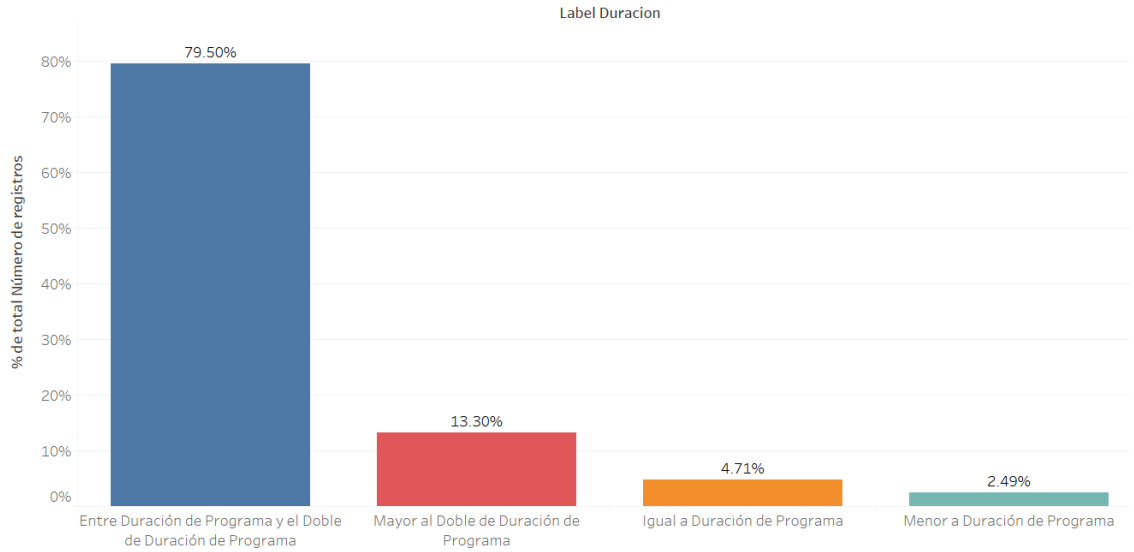


Figura 3-4: Distribución de las etiquetas de clasificación a nivel de Doctorado



3.3 Reglas de Asociación

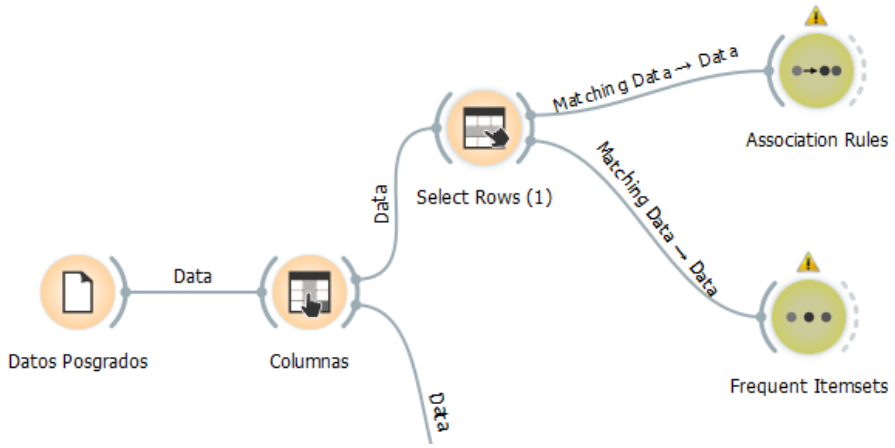
Una vez se tiene cada uno de los registros con su clasificación respectiva, se procedió a generar las reglas de asociación y el modelo. Para esto, se utilizó la herramienta de código abierto de aprendizaje automático (*Machine Learning*) y visualización de datos *Orange*[8]. Esta herramienta facilita el uso de los algoritmos de minería de datos, ya que se basa en el principio de programación visual. Lo que se hizo en *Orange*, en primera instancia, fue cargar los datos y ponerle un filtro de selección de columnas. Esto porque al empezar a hacer pruebas, y empezar a visualizar el modelo obtenido se observó que habían algunas variables que no le aportan al modelo, o que, por ejemplo, para hacer la predicción con los datos de estudiantes nuevos no se podían obtener algunas variables académicas como el P.A.P.A o el número de matrículas.

Una vez se tienen seleccionadas las columnas que se usarán para generar el modelo, se definió también un filtro de filas. Esto para lograr extraer reglas de asociación no tan comunes. Una vez se tienen estos 2 filtros listos, la herramienta permite obtener tanto los conjuntos más frecuentes

[8] *Orange, Data Mining, Fuitful and Fun,* Tomado de: https://orange.biolab.si/

como las reglas de asociación. Para este último, Orange utilizó el algoritmo *A-Priori*, una vez se tenía toda la secuencia programada (ver Figura 3-5) se podían empezar a tener en cuenta diferentes valores para los parámetros característicos del algoritmo *A-priori,* el soporte y la confianza. Usualmente, cuando se quieren obtener reglas interesantes dentro de un conjunto de datos se requiere que tanto el soporte como la confianza tomen valores lo más altamente posibles; pero esto no siempre se logra. Por ejemplo, para el conjunto de datos en este trabajo, lo primero que se hizo fue empezar a variar el parámetro de soporte, y lo que se observó fue que al existir tantas variables, y sobre todo al tener clasificados los registros en 4 grupos, no fue posible obtener reglas con un soporte mayor a 25%, esto cuando los 4 grupos tenían la misma cantidad de registros. Pero como en realidad no sucede de este modo, entonces se hicieron otras pruebas bajando el valor del soporte, hasta que se logró conseguir las primeras reglas para la clase *"Entre la duración del programa y el doble de la duración del programa"* con un soporte mínimo de 6%, y una confianza mínima de 75%.

Figura 3-5: Diagrama para obtener reglas de asociación en *Orange*



Para las reglas obtenidas para esta clase (2066 en total), se hizo una limpieza de las mismas y se pudieron obtener reglas que son interesantes de analizar y que pueden usarse como base para los programas que requieren generar un plan de mejoramiento. Entre estas reglas se pueden encontrar:

- {nivel=Maestría, area_conocimiento=Matemáticas y ciencias naturales} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.118, Conf: 0.79*).
- {nivel = Maestría, area_conocimiento = Matemáticas y ciencias naturales, tiposolicitud = Regular} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.117, Conf: 0.794*)
- {nivel = Maestría, area_conocimiento = Matemáticas y ciencias naturales, tiposolicitud = Regular, conveniointerins = No utilizan convenios} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.114, Conf: 0.793*)
- {facultad = Ciencias, nivel=Maestría} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.095, Conf: 0.79*)
- {nivel = Maestría, area_conocimiento = Matemáticas y ciencias naturales, estado_civil = Soltero } **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.078, Conf: 0.789*)
- {facultad = Ciencias, nivel= Maestría, estado_acreditacion = En proceso de autoevaluación } **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.075, Conf: 0.806*)
- {area_conocimiento = Matemáticas y ciencias naturales, tipo_plan_estudios = Profundización} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.071, Conf: 0.816*)
- {nbc = Educación} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.067, Conf: 0.75*)
- {programa = Maestría en Enseñanzas de las Ciencias Exactas y Naturales} **=>** Entre Duración de Programa y el Doble de duración de Programa (*Sop: 0.066, Conf: 0.828*)

En las reglas encontradas anteriormente, se puede observar obtener varias conclusiones. La primera es que (lo cual está acorde con lo observado en la Figura 3-3), si el nivel del programa de un registro es Maestría, se tiene una probabilidad de 0.688 (68.8% de los casos) de que el tiempo de duración sea "Entre la Duración del programa y el Doble de la duración". Además, en las reglas se puede ver como para las maestrías del área de Matemáticas y Ciencias Naturales, en un 79% de los casos, su duración está en el tercer grupo de duración. Ese 79% también se mantiene para los registros que pertenecen a una Maestría de la Facultad de Ciencias: Cabe anotar que esta facultad

no solo está en Bogotá, también hay una en la sede Medellín. Si la maestría tiene tipo plan de estudios de profundización y, además, pertenece al área de Matemáticas y Ciencias Naturales, entonces la probabilidad de que el tiempo de duración esté entre la duración y el doble de la duración del programa es de 0.816 (81.6% de los casos).

Para terminar, entre las reglas encontradas para esta etiqueta, también se observan 2 que están muy relacionadas entre sí: Si un graduado se graduó de un programa de pregrado perteneciente al NBC de Educación, hay una probabilidad de 0.75 (en un 75% de los casos) de que el registro pertenezca al tercer grupo, con respecto a la duración. Además, la otra regla expresa que si un graduado lo hizo del programa de Maestría en Enseñanzas de las Ciencias Exactas y Naturales, en un 82.8% de los casos, la duración se extendió más allá de lo reglamentado por el programa. Se menciona que las 2 reglas están relacionadas debido a que con el fondo que creó el Ministerio de Educación junto con Icetex, la ayuda que se ofrece, primero es para personas que se hayan graduado para ser docentes, y, en segundo lugar, la ayuda que se presta en específico para estudiar en la Universidad Nacional de Colombia, y para estudiar en el programa de Maestría en Ciencias Exactas y Naturales. Es decir, los que están siendo beneficiarios de esta ayuda del gobierno, están cumpliendo con sus estudios, pero gastan más tiempo del establecido por el programa.

Ahora, para analizar las otras etiquetas, se debía bajar el soporte aún más haciendo, que las reglas que se encontraran tuvieran soporte 1%. Entonces lo que se decidió fue tomar un subconjunto de los datos y hacer un análisis sobre este; en este subconjunto no se tuvieron en cuenta los registros que tenían clasificación “Entre la duración y el doble de la duración del programa”, por lo cual, el número de registros a tener en cuenta en este subconjunto fue de 10129, es decir, un 57.7% del conjunto original de datos. Para este subconjunto se estableció entonces un soporte mínimo de 30%, y se mantuvo un valor mínimo del 70% para la confianza. Por lo tanto, aplicando el filtro se encontraron las siguientes reglas de asociación para la etiqueta “Igual a Duración de Programa”:

- {nivel = Especialización} **=>** {label = Igual a Duración de Programa} (*Sop: 0.588, Conf: 0.777*)
- {lugarnacimiento = Nacional, tipo_plan_estudio = No Aplica} **=>** {label = Igual a Duración de Programa} (*Sop: 0.586, Conf: 0.778*)
- {tiposolicitud = Regular, tipo_plan_estudio = No Aplica} **=>** {label = Igual a Duración de Programa} (*Sop: 0.561, Conf: 0.834*)
- {estado_acreditación = En proceso de autoevaluación, tiposolicitud = Regular} **=>** {label = Igual a Duración de Programa} (*Sop: 0.458, Conf: 0.753*)
- {sede = Bogotá, estado_acreditacion = En proceso de autoevaluación, tiposolicitud = Regular, lugarnacimiento = Nacional} **=>** {label = Igual a Duración de Programa} (*Sop: 0.301, Conf: 0.899*)
- {nivel = Especialización, estado_acreditacion = En proceso de autoevaluación, tiposolicitud = Regular} **=>** {label = Igual a Duración de Programa} (*Sop: 0.400, Conf: 0.821*)
- {nivel = Especialización, tiposolicitud = Regular, estado_civil = Soltero} **=>** {label = Igual a Duración de Programa} (*Sop: 0.400, Conf: 0.821*)

En este conjunto de reglas se puede notar que el 77.7% de los estudiantes de las especializaciones se gradúan en el tiempo establecido por el programa. También se observa que casi todas las reglas de asociación contienen en el antecedente que el nivel del programa era especialización o que el tipo de plan de estudios era No Aplica, valor que solo está disponible para las especializaciones, ya que las maestrías son de investigación, profundización o ambas, y los doctorados son solo de investigación. El porcentaje de casos subió a 83.4%, cuando aparte de ser de especialización, el estudiante siguió un proceso de admisión regular. Todas estas asociaciones que incluyen a ese tipo de programas, se podrían explicar por el hecho de que existen algunos programas de especialización en la Universidad Nacional de Colombia que no exigen la realización de un trabajo final, solamente cumplir con un número definido de asignaturas para poder obtener el título. Entonces de ahí puede surgir la razón para que el 77.7% de los graduados de una especialización completen sus estudios en 2 o 3 semestres, dependiendo de los programas particulares.

Otra regla interesante se presenta cuando un graduado ingresó por admisión regular, y también se graduó de un programa que está en proceso de autoevaluación. Cuando estas 2 condiciones ocurren, entonces en el 75.3% de los casos pertenecía a la clase Igual a la duración del programa. Si a esas 2 condiciones se le suma que sea un programa de la sede Bogotá, también que cuando el graduado ingresó al programa estaba soltero y que fuera colombiano, entonces esa probabilidad fue de 0.899 (89.9% de los casos).

Para las 2 clases faltantes y para no bajar tanto el soporte en el subconjunto anterior, se decidió hacer un nuevo subconjunto en el cual solo se tenían en cuenta las clases "Menor al tiempo de duración" y "Mayor al doble del tiempo de duración". De ese modo, se obtienen 3370 registros, o un 19.15% del conjunto de datos original. Entonces, teniendo en cuenta estas cifras se estableció que el soporte mínimo fuera de 20% y la confianza mínima de 75%. Entonces, se obtuvieron como reglas más relevantes las siguientes:

- {nivel = Especialización, universidadpregrado = Universidad Nacional de Colombia - UN} **=>** {label = Menor a Duración de Programa} (*Sop: 0.326, Conf: 0.972*)
- {nivel = Especialización, lugarnacimiento = Nacional, estado_civil = soltero, conveniointerins = No utilizan convenios} **=>** {label = Menor a Duración de Programa} (*Sop: 0.388, Conf: 0.958*)
- {nivel = Especialización, estado_civil = Soltero, conveniointerins = No utilizan convenios} **=>** {label = Menor a Duración de Programa} (*Sop: 0.388, Conf: 0.958*)
- {nivel = Especialización, genero = F} **=>** {label = Menor a Duración de Programa} (*Sop: 0.211, Conf: 0.931*)

- {tipo_plan_estudios = Investigación, nivel = Maestría} **=>** {label = Mayor al Doble de Duración de Programa} (*Sop: 0.304, Conf: 0.957*)
- {tipo_plan_estudios = Investigación } **=>** {label = Mayor al Doble de Duración de Programa} (*Sop: 0.318, Conf: 0.951*)
- {sede => Bogotá, nivel = Maestría} **=>** {label = Mayor al Doble de Duración de Programa} (*Sop: 0.324, Conf: 0.945*)

- { nivel = Maestría} **=>** {label = Mayor al Doble de Duración de Programa} (*Sop: 0.436, Conf: 0.913*)
- { nivel = Maestría, tipos_solicitud = Regular} **=>** {label = Mayor al Doble de Duración de Programa} (*Sop: 0.297, Conf: 0.944*)
- { nivel = Maestría, genero = M} **=>** {label = Mayor al Doble de Duración de Programa} (*Sop: 0.250, Conf: 0.909*)

Teniendo en cuenta que estas reglas sólo tienen en cuenta el 19.15% de los datos, se puede concluir que en el 97.2% de los casos, una persona que haga una especialización y haya obtenido el título de pregrado de la Universidad Nacional, termina el programa en menos tiempo; esto se debe a una modalidad implementada en muchas facultades, en la cual es permitido que un estudiante pueda tomar asignaturas de posgrado mientras estudia su pregrado, y que una vez sea admitido al programa, pueda homologar esas asignaturas. De este modo, para muchos estudiantes, el último semestre de pregrado se convierte en el primer semestre "no oficial" del programa de posgrado, como en este caso y de acuerdo con los datos de los programas de especialización. Con base en esta regla se generaron otras reglas que incluyen más campos; por ejemplo: si el nivel del programa es especialización, la persona tiene nacionalidad colombiana, no cursó su programa con un convenio y al iniciar el programa es soltero, entonces en un 95.8% de los casos, el tiempo de duración fue menor al tiempo establecido por el programa.

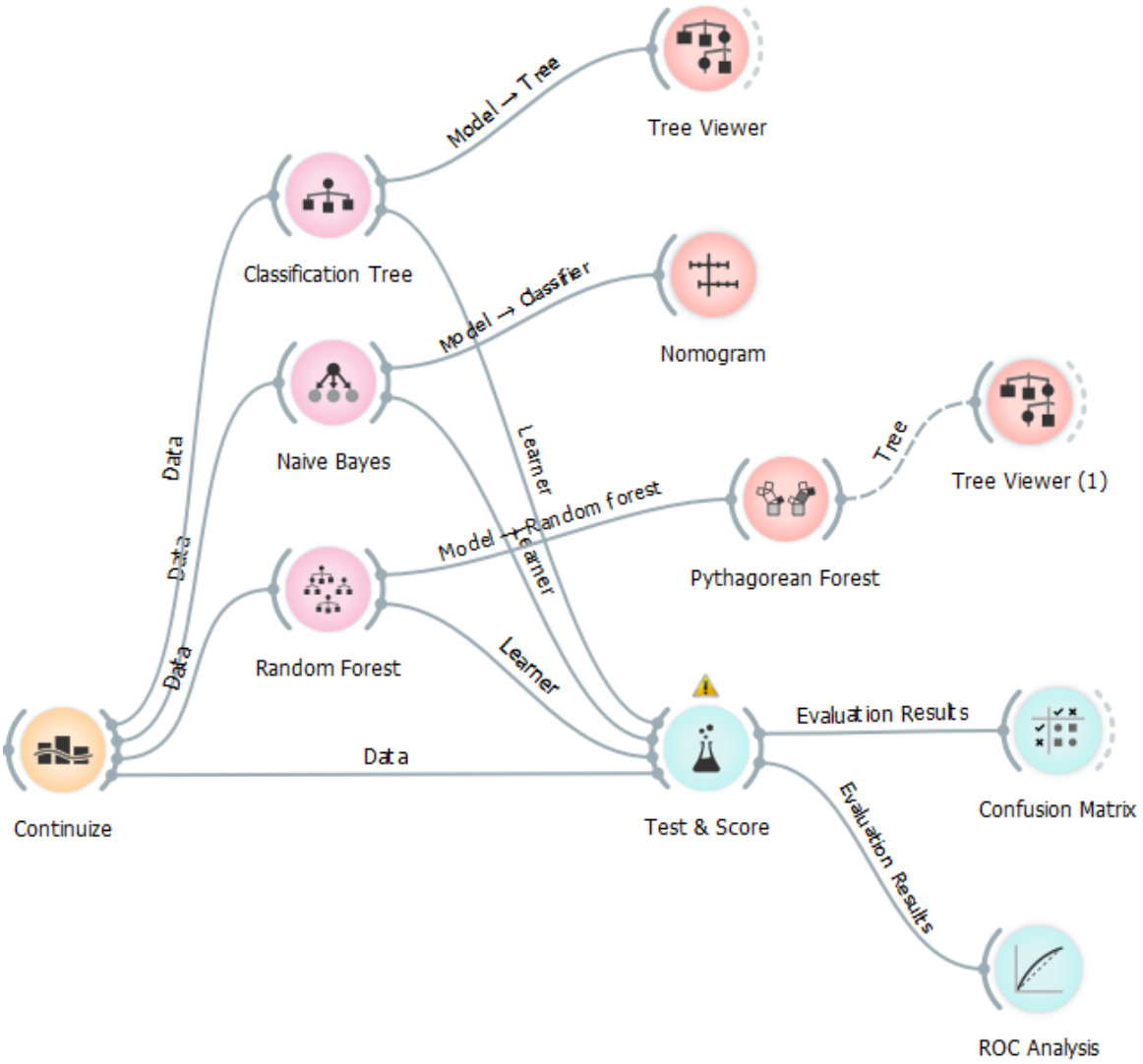
Para la clase Mayor al Doble de Duración de Programa, también se encontraron reglas interesantes. Por ejemplo, de los registros analizados, en un 95.7% de los casos, una persona se graduó en más del doble del tiempo de duración definido por el programa cuando el programa al que pertenecía era una maestría y el plan de estudios era de Investigación. Esta es una regla interesante que se puede explicar por el hecho de que los programas de investigación de maestría o de doctorado tienen una mayor incertidumbre (debido al trabajo autónomo durante el desarrollo de las tesis) que las maestrías de profundización y las especializaciones. Si aparte de ser de maestría, este programa es de la sede Bogotá, se observa que en un 94.5% de los casos la duración del programa fue mayor al doble. Vale resaltar, que para este análisis se tuvieron en cuenta solamente el 19.15% de los datos. Algo interesante observado en estos datos es que si un registro cumple con ser de

maestría y ser hombre, entonces en el 90.9% de los casos, el tiempo de duración fue mayor al doble del tiempo reglamentado por el programa.

3.4 Generación del Modelo

Otra parte importante de este trabajo fue la generación de un modelo que permita hacer la clasificación o predicción para nuevos estudiantes que hasta ahora fueron admitidos, o que se encuentran en el primer semestre. Esto se hizo con el apoyo del software Orange (ver Figura 3-6). Lo primero que se hizo fue convertir las variables discretas a valores binarios, es decir cada dimensión con sus posibles valores se convirtió en una nueva dimensión con solo 2 valores posibles 0 y 1. Después de tener los datos con esta distribución binaria de los datos, se procedió a utilizar 3 algoritmos de clasificación: Naive Bayes, árbol de decisión y bosque aleatorio (*Random Forest*).

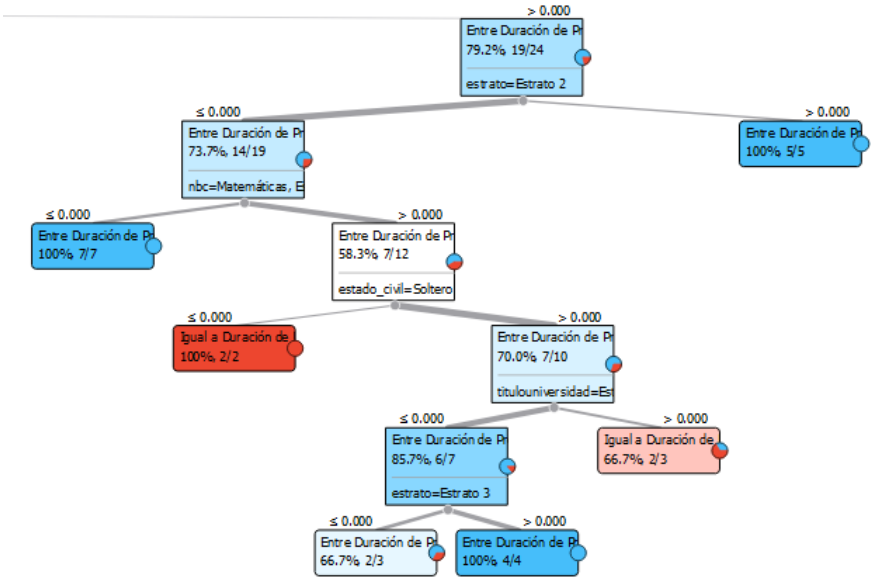
Figura 3-6: Diagrama del modelo de clasificación en Orange



Cada uno de los algoritmos se aplicó para generar los resultados sobre el conjunto de datos de entrada. En el caso del árbol de decisión, se configuró el programa para que fuera de solo 2 ramas por cada selección que se hiciera en el registro, para que al final y después de ir recorriendo cada una de las ramas el algoritmo lo clasificará en una de las 4 clases disponibles. En la figura 3-7 se muestra el árbol de decisión. Allí se ve por ejemplo que si el registro en la variable Estrato tiene el valor Estrato 2, entonces esa rama acaba ahí y se le asigna la clase “Entre la duración y el doble de la duración del programa”. Si por el contrario, no es Estrato 2, entonces debe seguir revisando con cada una de los valores de las siguientes variables hasta poder llegar a una clasificación particular. Uno de los problemas que tiene este algoritmo, como en esta investigación, es que al tener tantas

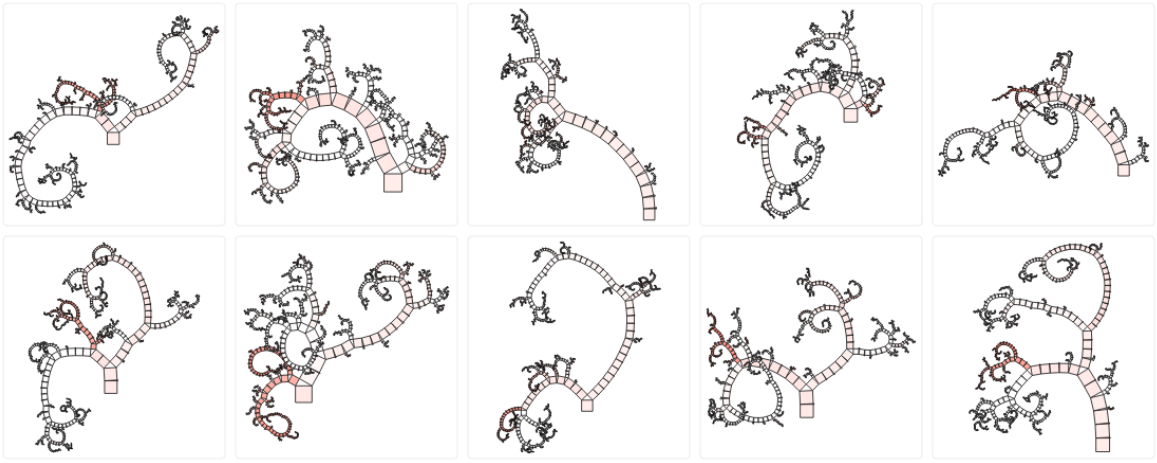
variables disponibles, así como etiquetas de clasificación disponibles, los árboles de decisión tienen una gran profundidad y anchura. Entonces se hacen un poco difíciles de leer, y no muy útiles a la hora de presentar los resultados concretos a los programas particulares del Sistema de Autoevaluación.

Figura 3-7: Ejemplo de visualización de un árbol de decisión



Otro algoritmo que se utilizó para el modelo de clasificación fue el de bosque aleatorio. Para este algoritmo se definieron los parámetros para que se generarán 10 árboles de decisión. De esta manera, y como se puede observar en la Figura 3-8, se generaron árboles que por su forma son diferentes y que se pueden filtrar por cada una de las clases. En *Orange*, cuando ya se tiene seleccionado uno de los árboles generados, se puede revisar en el componente visualizador de árbol, y se mostrará dicho árbol con todas sus ramas (similar al que se muestra en la Figura 3-7). Al igual que un árbol sencillo de decisión, la problemática de visualización de los resultados es una dificultad que se presenta en este caso, teniendo en mente a los usuarios finales del producto de esta investigación. Figuras como la 3-7 y 3-8, no brindan a los programas, en principio, mucha información y también pueden ser difíciles de leer.

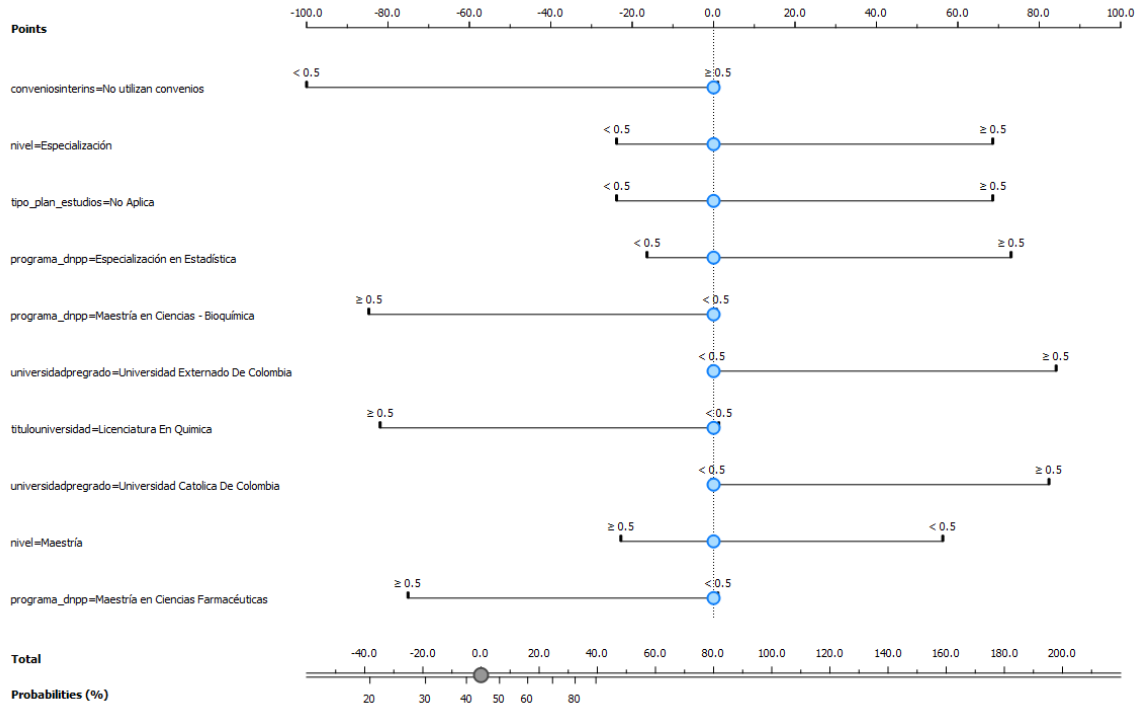
Figura 3-8: Visualización de los arboles obtenidos del *Random Forest*



Elaboración Propia

Finalmente, el último algoritmo de clasificación usado fue Naive Bayes. Como ya se ha explicado, este algoritmo no genera un árbol, pero si genera unos resultados que también se pueden presentar de manera visual. Para este fin, *Orange* presentan un módulo llamado Nomograma (ver Figura 3-9) y en él se pueden ver cuáles son las variables que más afectan al modelo, así cada clase se asigna una probabilidad particular. En la Figura 3-9, por ejemplo, se muestran los valores que deberían tener las 10 variables con más influencia, en este caso, sobre la etiqueta “Igual a la duración del programa”. Así se puede también visualizar si una variable tiene influencia positiva o negativa para la clase. El nomograma se ve como una visualización de los modelos y puede ser de utilidad y fácil entendimiento para los representantes de los programas de posgrado.

Figura 3-9: Visualización de los resultados del algoritmo Naive Bayes por medio de un Nomograma



Cómo se puede observar, los 3 algoritmos de clasificación utilizados muestran modelos y visualizaciones diferentes para presentar los resultados. La visualización que permite más facilidad de lectura y entender mejor los resultados es la que provee el algoritmo de *Naive Bayes*, claro, todo esto pensando en el módulo que estará en el Sistema de Autoevaluación. Ahora, lo que se debe analizar cuál de los 3 modelos es el mejor. Para esto se debe hacer la evaluación de los tres algoritmos utilizados.

3.5 Evaluación del Modelo

Para hacer la evaluación de los diferentes algoritmos, lo que se hizo fue entrenar a los algoritmos, para después hacer que cada uno tomará una parte de los datos para que pudiera hacer pruebas, para realizar este procedimiento se utilizó un módulo de *Orange* utilizado para probar este tipo de

modelos (Ver Figura 3-6) y allí se hizo uso de un método llamado Validación cruzada, en el cual se toma un 90% de los datos para entrenar el modelo, y al 10% restante se le aplican las pruebas del modelo, en este caso se parametrizó para que este procedimiento se repitiera 10 veces. Así cada algoritmo le da una etiqueta a cada registro y se compara con la etiqueta original, para determinar cuántos registros están bien clasificados. Para revisar éstas medidas y con ayuda del software Orange, se generaron las matrices de confusión y las curvas ROC respectivas, las matrices de confusión muestran 17598 registros de prueba ya que se acumulan los resultados obtenidos en cada uno de los 10 experimentos hechos con validación cruzada.

3.5.1 Matrices de confusión

Para las siguientes matrices de confusión se calculan cuantas predicciones son correctas y cuantas no para cada una de las clases, con cada uno de los algoritmos: Naive Bayes (Ver Figura 3-10), Random Forest (Ver Figura 3-11) y Árboles de decisión (Ver figura 3-12). A partir de estas matrices se pueden obtener 2 variables que son importantes para definir si un algoritmo se puede considerar como buen clasificador o no, esas variables son exactitud y tasa de error:

Figura 3-10: Matriz de confusión para el algoritmo *Naive Bayes*

Predicted

Actual	Entre Duraci...	Igual a Duraci...	Mayor al Doble ...	Menor a Duraci...	Σ
Entre Duraci...	74.0 %	23.1 %	61.2 %	14.3 %	7469
Igual a Duraci...	13.1 %	68.9 %	6.3 %	48.5 %	6759
Mayor al Doble ...	10.7 %	1.3 %	31.3 %	0.7 %	1615
Menor a Duraci...	2.3 %	6.7 %	1.2 %	36.5 %	1755
Σ	4882	6317	3170	3229	17598

Elaboración Propia

Predicted

Actual	Entre Duraci...	Igual a Duraci...	Mayor al Doble ...	Menor a Duraci...	Σ
Entre Duraci...	3611	1457	1940	461	7469
Igual a Duraci...	638	4354	201	1566	6759
Mayor al Doble ...	521	81	991	22	1615
Menor a Duraci...	112	425	38	1180	1755
Σ	4882	6317	3170	3229	17598

Elaboración Propia

Para el algoritmo *Naive Bayes*, se obtiene la matriz de confusión que se ve en la Figura 3-10, a partir de ellas se puede obtener los siguientes valores para las medidas de desempeño:

$$Exactitud = \frac{Número\ de\ Predicciones\ Correctas}{Total\ de\ Registros} = \frac{3611 + 4354 + 991 + 1180}{17598}$$

$$Exactitud = 0.5759 = 57.6\ \%$$

$$Tasa\ de\ error = \frac{Número\ de\ Predicciones\ Incorrectas}{Total\ de\ registros} = 1 - Exactitud$$

$$Tasa\ de\ error = 1 - 0.5759 = 0.4241 = 42.41\%$$

Figura 3-11: Matriz de confusión para el algoritmo *Random Forest*

Predicted

Actual	Entre Duraci...	Igual a Duraci...	Mayor al Doble ...	Menor a Duraci...	Σ
Entre Duraci...	69.3 %	19.4 %	54.7 %	5.4 %	7469
Igual a Duraci...	13.5 %	71.0 %	3.6 %	30.2 %	6759
Mayor al Doble ...	15.4 %	1.1 %	41.0 %	0.5 %	1615
Menor a Duraci...	1.9 %	8.6 %	0.7 %	64.0 %	1755
Σ	8049	7313	717	1519	17598

Elaboración Propia

Predicted

Actual	Entre Duraci...	Igual a Duraci...	Mayor al Doble ...	Menor a Duraci...	Σ
Entre Duraci...	5577	1418	392	82	7469
Igual a Duraci...	1086	5189	26	458	6759
Mayor al Doble ...	1236	78	294	7	1615
Menor a Duraci...	150	628	5	972	1755
Σ	8049	7313	717	1519	17598

Elaboración Propia

Solo observando la matriz de confusión, el algoritmo *Random Forest* (ver figura 3-11), parece ser un mejor clasificador que el algoritmo *Naive Bayes*. Pero para confirmar esta observación es necesario calcular las mismas medidas:

$$Exactitud = \frac{Número\ de\ Predicciones\ Correctas}{Total\ de\ Registros} = \frac{5577 + 5189 + 294 + 972}{17598}$$

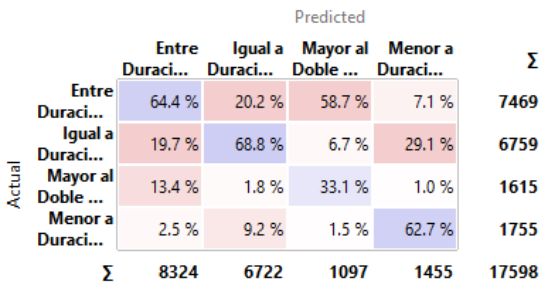
$$Exactitud = 0.6837 = 68.37\ \%$$

$$Tasa\ de\ error = \frac{Número\ de\ Predicciones\ Incorrectas}{Total\ de\ registros} = 1 - Exactitud$$

$$Tasa\ de\ error = 1 - 0.6837 = 0.3163 = 31.63\%$$

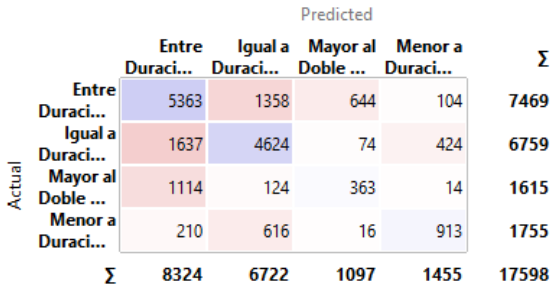
Con estos resultados se confirma que el algoritmo *Random Forest* supera al algoritmo Naive Bayes, obteniendo un 31.63% como tasa de error, que da una posibilidad de obtener una mejor predicción para registros de nuevos estudiantes.

Figura 3-12: Matriz de confusión para el algoritmo de árbol de decisión



Actual \ Predicted	Entre Duraci...	Igual a Duraci...	Mayor al Doble ...	Menor a Duraci...	Σ
Entre Duraci...	64.4 %	20.2 %	58.7 %	7.1 %	7469
Igual a Duraci...	19.7 %	68.8 %	6.7 %	29.1 %	6759
Mayor al Doble ...	13.4 %	1.8 %	33.1 %	1.0 %	1615
Menor a Duraci...	2.5 %	9.2 %	1.5 %	62.7 %	1755
Σ	8324	6722	1097	1455	17598

Elaboración Propia



Actual \ Predicted	Entre Duraci...	Igual a Duraci...	Mayor al Doble ...	Menor a Duraci...	Σ
Entre Duraci...	5363	1358	644	104	7469
Igual a Duraci...	1637	4624	74	424	6759
Mayor al Doble ...	1114	124	363	14	1615
Menor a Duraci...	210	616	16	913	1755
Σ	8324	6722	1097	1455	17598

Elaboración Propia

Ahora, el algoritmo de árbol de decisión al tener un funcionamiento parecido al del algoritmo *Random Forest*, se espera que obtenga resultados parecidos. A simple vista, la matriz de confusión, tiene cifras similares a las obtenidas para el algoritmo del bosque aleatorio:

$$Exactitud = \frac{Número\ de\ Predicciones\ Correctas}{Total\ de\ Registros} = \frac{5363 + 4624 + 363 + 913}{17598}$$

$$Exactitud = 0.640 = 64\ \%$$

$$Tasa\ de\ error = \frac{Número\ de\ Predicciones\ Incorrectas}{Total\ de\ registros} = 1 - Exactitud$$

$$Tasa\ de\ error = 1 - 0.640 = 0.36 = 36\%$$

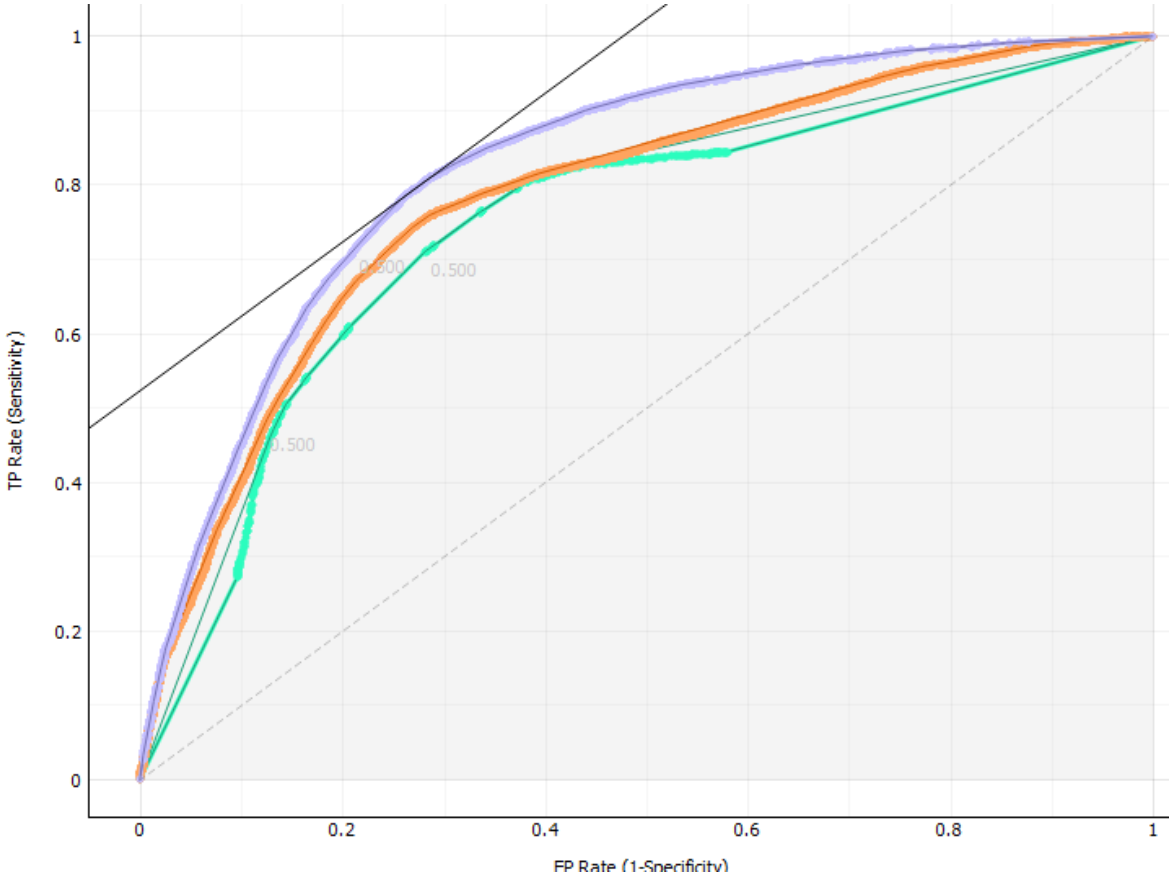
Entonces, teniendo los valores de exactitud y de tasa de error, se observa que los 2 algoritmos que tienen como base árboles de decisión tienen mejor rendimiento que *Naive Bayes*. Luego para la posterior predicción que se va a realizar de estudiantes que hasta ahora ingresaron a un programa de posgrado en el año 2017, lo más confiable será tener en cuenta la clasificación propuesta para cada registro por el método basado en bosque aleatorio.

3.5.2 Curvas ROC

Aparte de las medidas de rendimiento anteriormente expuestas, para los algoritmos de clasificación, y con ayuda también de las matrices de confusión, se pueden generar curvas ROC, que son curvas en las cuales se puede tener una gráfica de la tasa de falsos positivos en comparación con la tasa de verdaderos positivos. Para poder obtener estos valores, es necesario tener matrices independientes para cada una de las clases, entonces el software Orange, permite obtener las gráficas a partir de las matrices de confusión mostradas anteriormente.

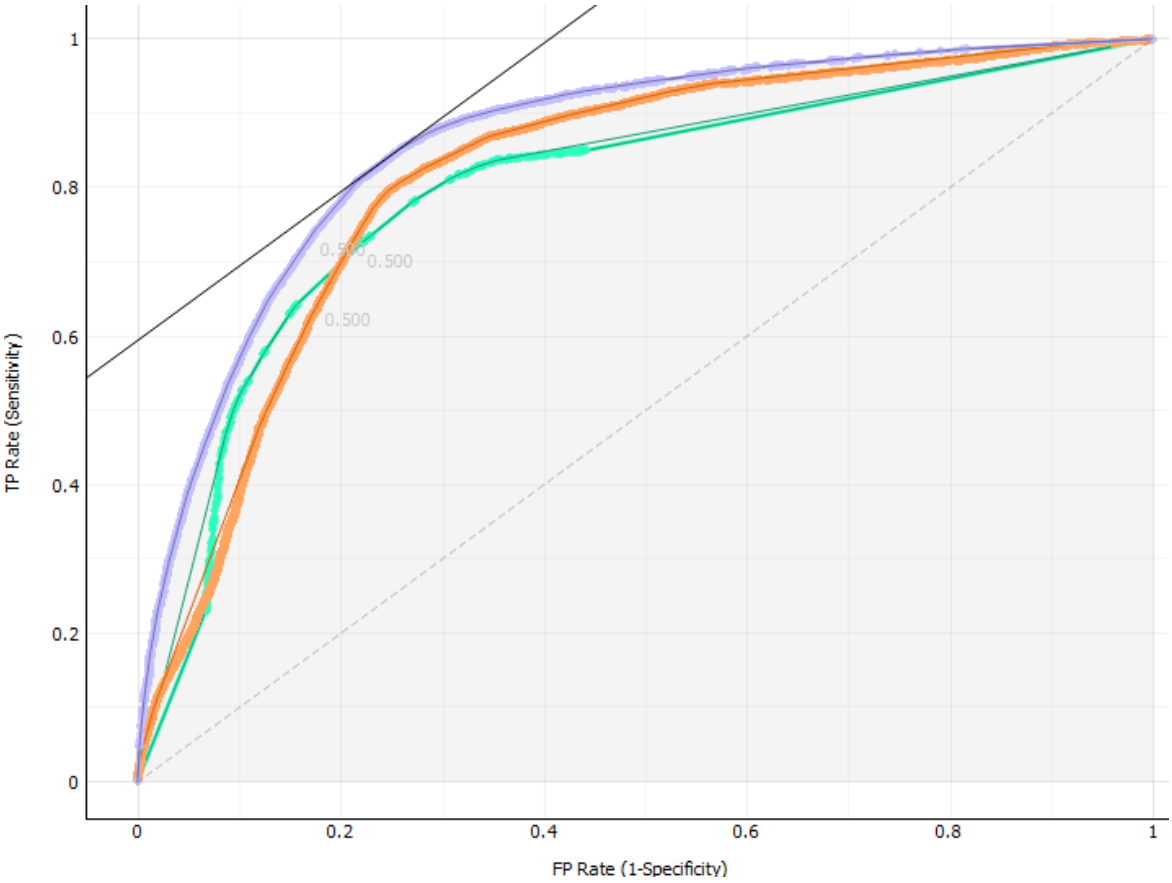
Para la clase "Entre duración del programa y el doble de duración de programa", se puede observar en la Figura 3-13 que los 3 algoritmos tienen un comportamiento similar, teniendo al Random Forest (Morado) como el algoritmo de mejor comportamiento, y el Naive Bayes (naranja) como el segundo mejor superando en casi toda la gráfica a la curva ROC producida por el algoritmo de Árbol de Decisión (Verde), aunque los 3 algoritmos están bastante alejados de tener un área bajo la curva de 1 que sería el clasificador ideal.

Figura 3-13: Curva ROC para la clase “Entre duración de programa y el doble de duración de programa” - Random Forest (Morado), Naive Bayes (naranja) y Árbol de Decisión (Verde)



Para la clase "Igual a duración de programa" (ver Figura 3-14), las curvas ROC de los 3 algoritmos tienen un comportamiento similar, aunque se puede observar que para este caso el algoritmo árbol de decisión tiene mejor comportamiento que el *Naive Bayes* en la primera mitad de la curva, sin embargo se puede notar como los 3 algoritmos tienen unos resultados bastante similares. Sin embargo, *Random Forest* se mantiene como el mejor clasificador.

Figura 3-14: Curva ROC para la clase “Igual a duración de programa” - Random Forest (Morado), Naive Bayes (naranja) y Árbol de Decisión (Verde)



Para las clases "Mayor al doble de duración de programa" (ver figura 3-15) y "Menor a duración de programa" (ver Figura 3-16), la tendencia de las curvas cambia un poco, para la primera clase, el algoritmo con mejor comportamiento es *Naive Bayes,* aunque con el algoritmo *Random Forest* se mantiene cercano, en cambio el algoritmo de árbol de decisión tiene un comportamiento bastante malo en comparación a los otros 2 algoritmos. Para la clase “Menor a duración de programa” se vuelve a tener a Random Forest como el algoritmo con mejor comportamiento, y los otros 2 algoritmos con un comportamiento similar se mantienen ambos en segundo lugar.

Figura 3-15: Curva ROC para la clase “Mayor al doble de duración de programa” - Random Forest (Morado), Naive Bayes (naranja) y Árbol de Decisión (Verde)

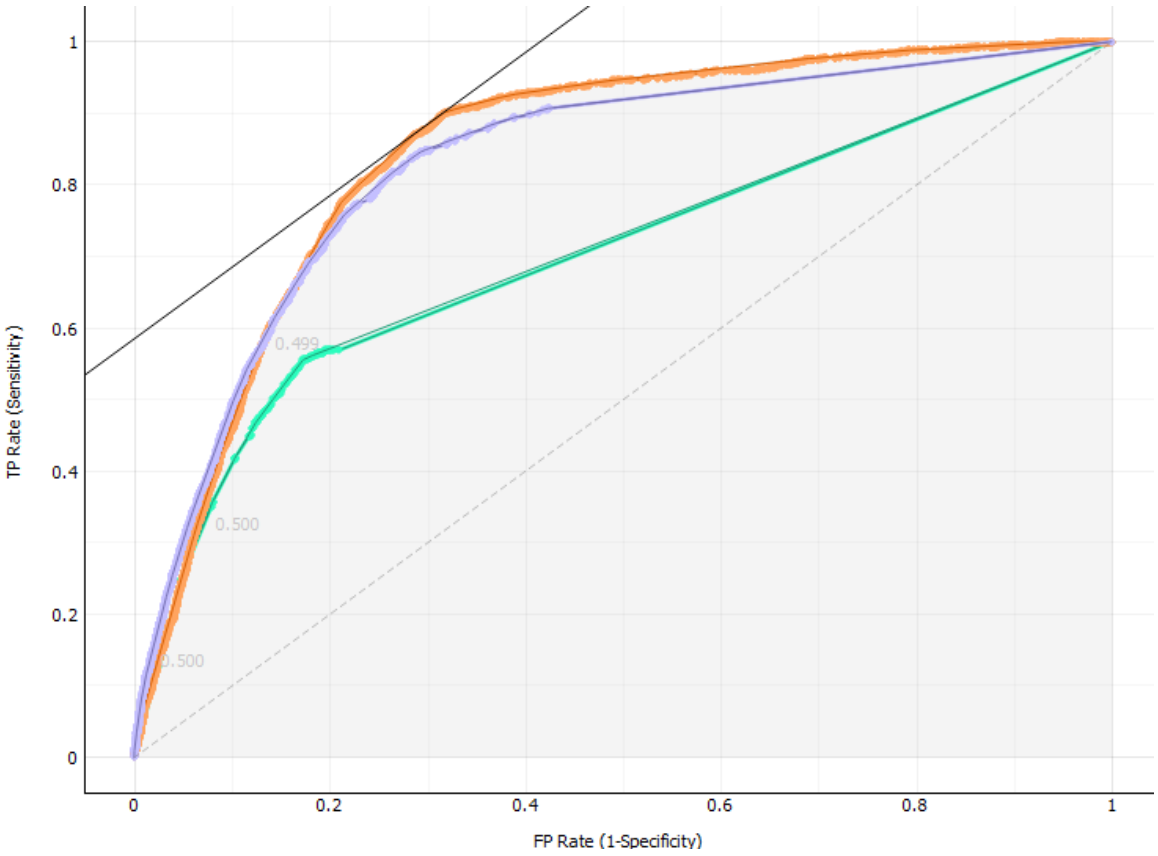
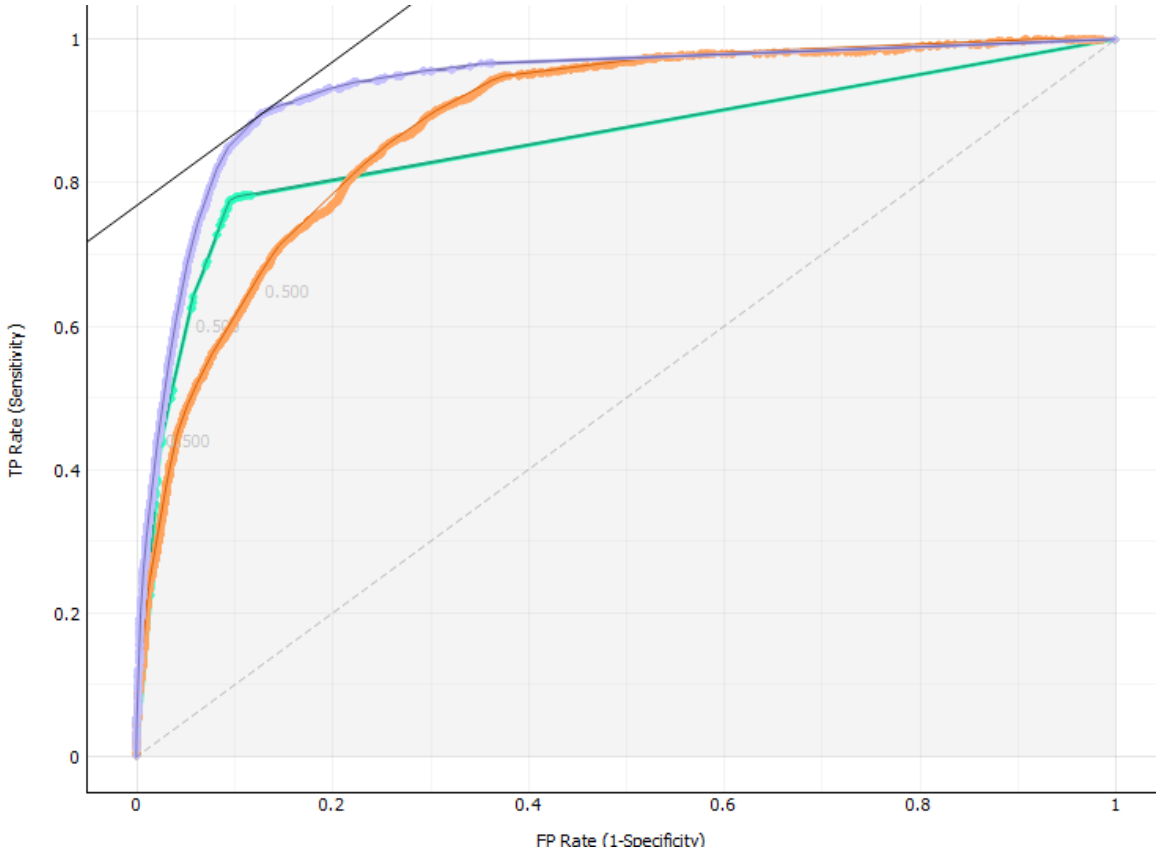


Figura 3-16: curva ROC para la clase "Menor a duración de programa" - Random Forest (Morado), Naive Bayes (naranja) y Árbol de Decisión (Verde)



En conclusión, el mejor algoritmo para hacer la clasificación que se requería para esta investigación fue el algoritmo basado en *Random Forest*, que tuvo medida de exactitud alta, y como se puede observar en la Figura 3-17 tiene mejores medidas de rendimiento que los otros 2 algoritmos. Los otros 2 algoritmos también tienen buenos resultados en cuanto a sus medidas de rendimiento, pero también con las curvas ROC, en 3 de ellas las 2 curvas ROC estuvieron muy cercanas entre sí, y también se mantuvieron muy cercanas a los puntos altos del algoritmo *Random Forest.* Por lo cual, a pesar de estar tan cercanos en cuanto a resultados, los algoritmos de Random Forest y árbol de decisión si obtuvieron una medida de exactitud mucho más alta que Naive Bayes, por tal motivo, se podría también utilizar el algoritmo basado en árboles de decisión para hacer predicciones.

Figura 3-17: Medidas de rendimiento para los algoritmos de clasificación utilizados en el modelo

Method	AUC	CA	F1	Precision	Recall
Random Forest	0.858	0.684	0.737	0.710	0.768
Naive Bayes	0.812	0.576	0.666	0.689	0.644
Tree	0.798	0.640	0.686	0.688	0.684

4. Predicción de casos nuevos y adecuación en el Sistema de Autoevaluación

En este capítulo se presentan las aplicaciones que se hicieron con el modelo generado en capítulos anteriores, en donde, en primer lugar, se aplicó un modelo predictivo para estudiantes nuevos, es decir estudiantes que fueron admitidos en el año 2016 para algún programa de posgrado, y en segundo lugar se generó un módulo en el Sistema de Autoevaluación para que los programas pudieran visualizar en el módulo de plan de mejoramiento los resultados tanto de las asociaciones, como de la predicción para nuevos estudiantes, y esto con el fin de que puedan sacar algunas conclusiones, y puedan generar acciones que permitan mejorar la situación de los estudiantes tanto actualmente, como en un futuro.

4.1 Predicción de casos nuevos

Una vez se tuvo el modelo generado y evaluado, se procedió a conseguir la información de los admitidos que tuvieron los programas de posgrado en el año 2017, para ambos semestres, logrando conseguir 4750 registros.

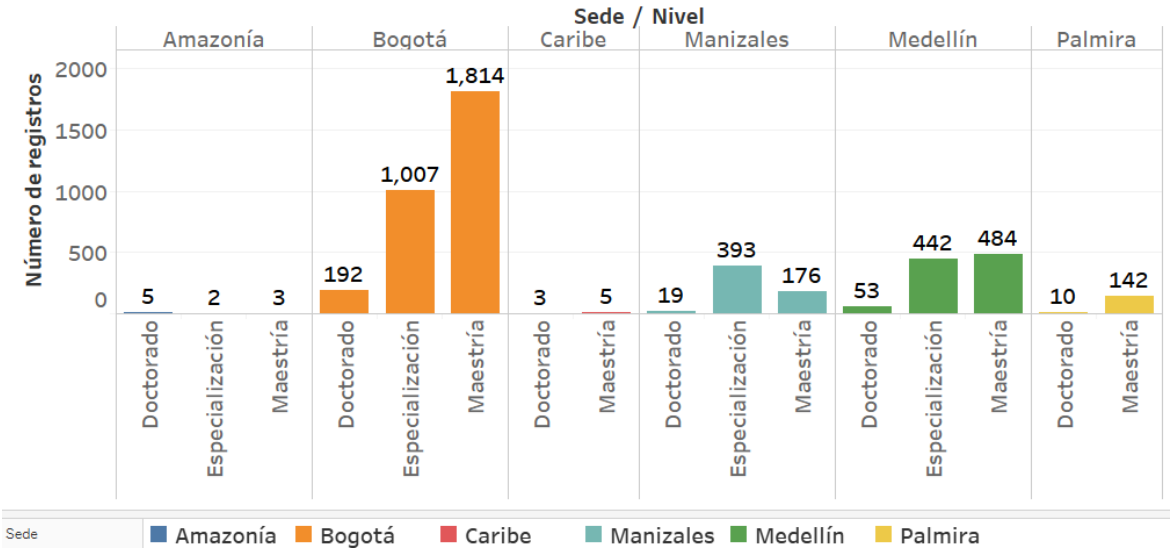
Para todos estos registros se hizo el mismo proceso de limpieza de datos que se hizo con el conjunto de datos de graduados, es decir, se limpiaron registros duplicados, se eliminaron variables que no aportaban mucho al modelo. De igual manera, las variables debían ser las mismas que se tuvieron en cuenta para el modelo, con los mismos nombres y las mismas opciones para sus valores.

También, se realizó la limpieza de los campos titulo_pregrado y universidad_pregrado, para que, al igual que se hizo con los graduados, se mantuvieran valores iguales para los diferentes tipos de

títulos y universidades que pudieran presentar similitud. Los campos NBC y categoria_edad también fueron creados y calculados, manteniendo los mismos parámetros que se tuvieron en cuenta para los datos de las personas que se graduaron.

Al hacer esta limpieza de datos y pre-procesamiento se pudieron generar los siguientes gráficos que permiten tener una vista general de los datos obtenidos. En la Figura 4-1 se puede observar como la tendencia de estos estudiantes admitidos es que la mayoría se dieron en la sede Bogotá, y que con una gran diferencia, la mayoría ingresaron a un programa de Maestría. En las otras sedes, las cifras son mucho más bajas y la diferencia no es tan notoria; por ejemplo, en la sede Medellín ambos niveles son muy parejos, al igual que en Amazonía, a pesar de la baja cantidad de datos. La sede Manizales, por el contrario, tiene muchos más admitidos para especializaciones que para maestría; y en todas las sedes, el número de admitidos para programas de Doctorado es muy bajo, no pasan de 300 admitidos en total.

Figura 4-1: Distribución de los datos de los admitidos divididos por sede y nivel



Cuando se realizó el análisis descriptivo de los graduados en el capítulo 2, se observaron datos interesantes; por ejemplo, el que tenía que ver con convenios, donde se observaba que la Facultad de Derecho en Bogotá se llevaba la gran mayoría del número de convenios que se tenían a nivel nacional, entonces para este conjunto de datos también se quiso revisar esa variable de convenios. Se pudo observar (como se muestra en la Figura 4-2) que la mayoría de los convenios están en las

especializaciones en la sede Bogotá, y, más específicamente, todos estos convenios son de Derecho (ver Figura 4-3). Luego esta variable se sigue comportando de manera similar, lo que hace pensar que para la predicción de estos casos, su tiempo de permanencia podría ser Igual a la duración del programa.

Figura 4-2: Distribución de los admitidos por sede y nivel, clasificándolo por el uso de convenios

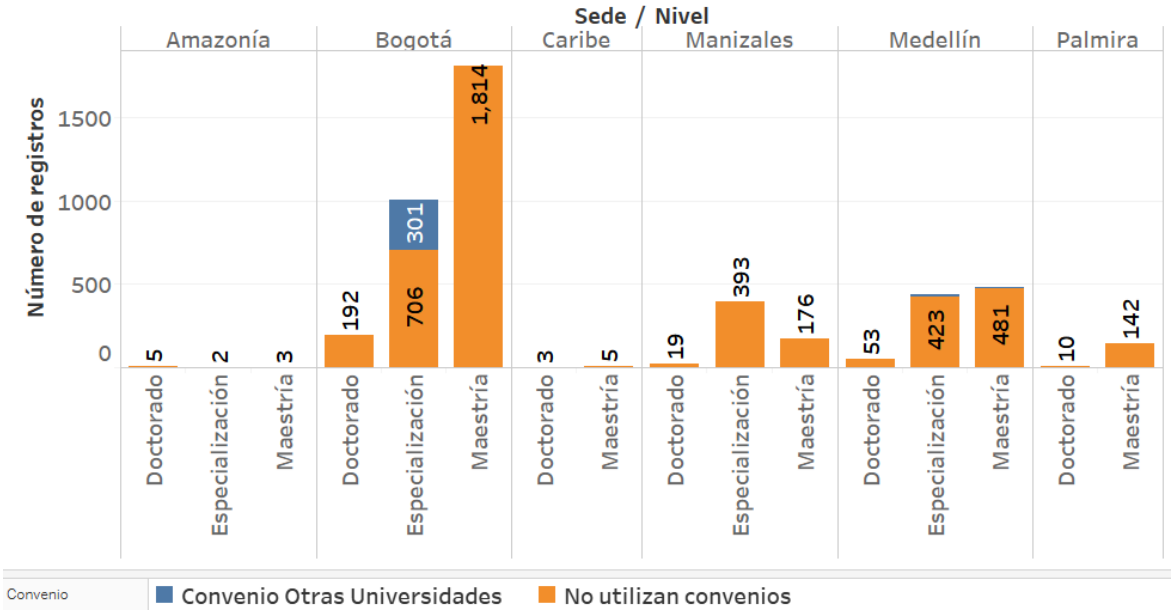
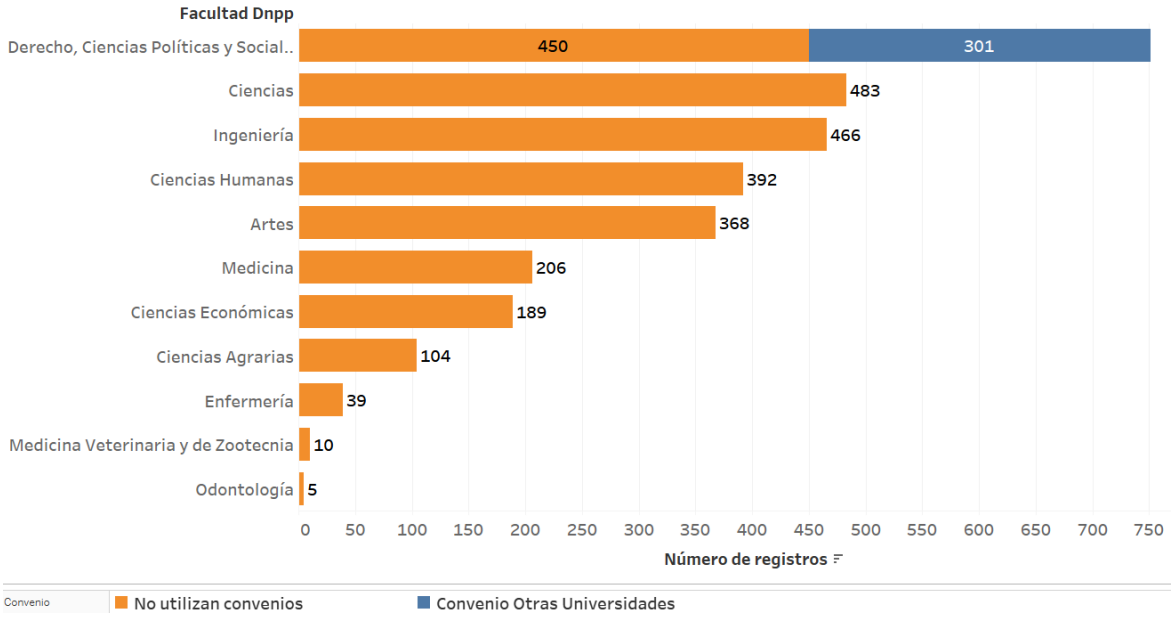
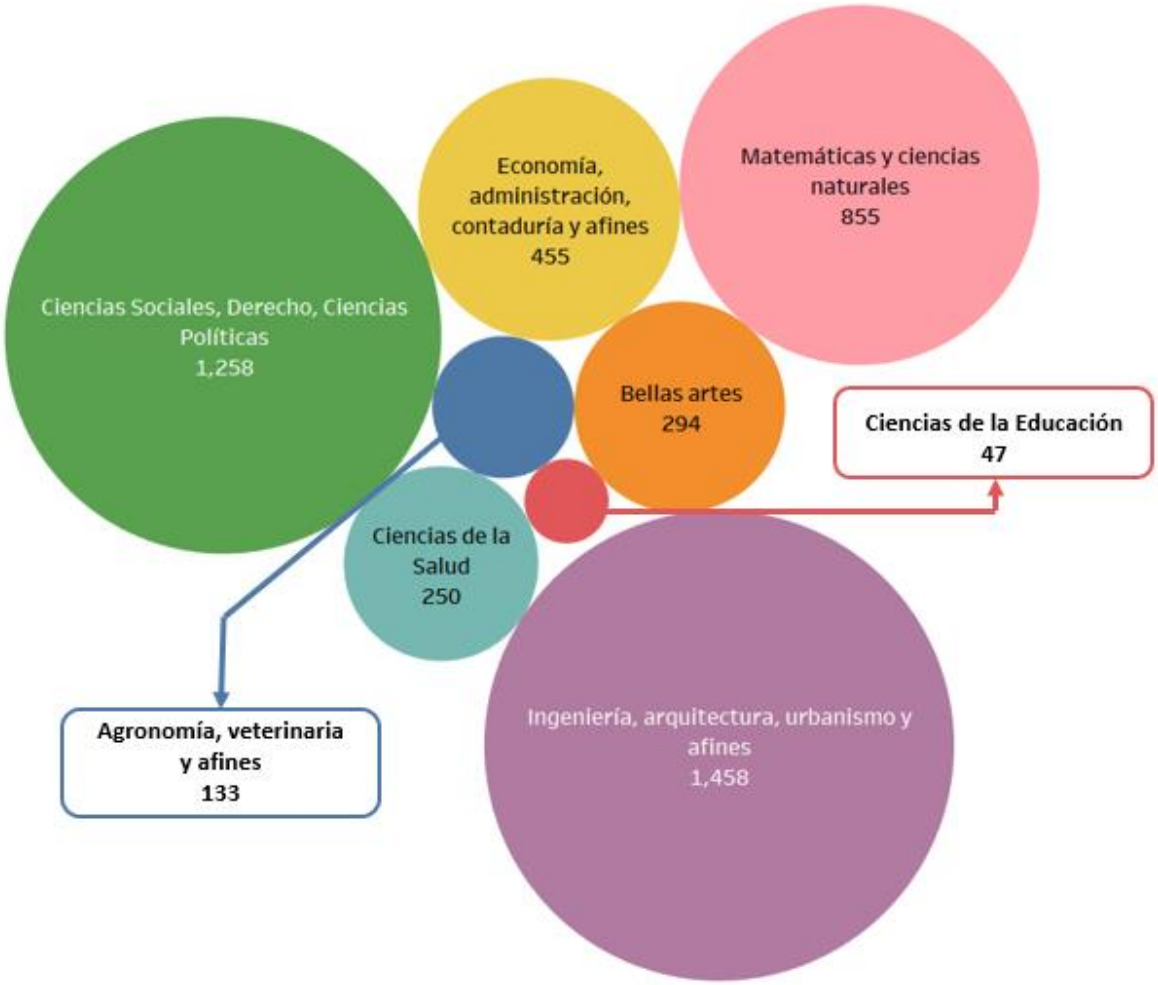


Figura 4-3: Proporción de convenios para los admitidos de la sede Bogotá



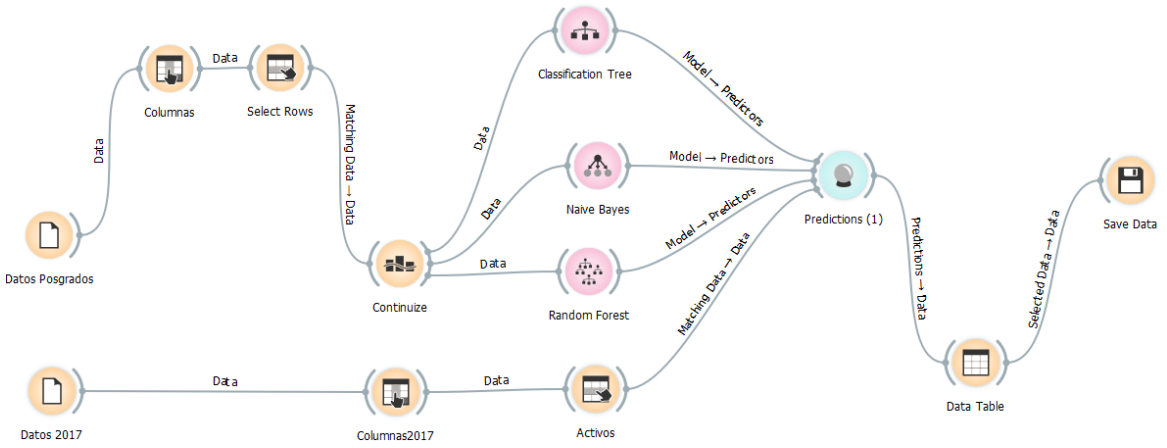
Otro análisis que se quería hacer de los datos de los matriculados era identificar el área de conocimiento con mayor cantidad de admitidos (ver Figura 4-4), y hacer la comparación con los datos de graduados. Entonces, observando la misma gráfica para los graduados (ver Figura 2-7), se pudo ver que el patrón se mantuvo, a pesar de que el área de conocimiento con mayor cantidad de admitidos fue Ingeniería, arquitectura, urbanismo y afines; el área de conocimiento de Ciencias Sociales, Derecho y Ciencias Políticas se mantiene muy cerca, siendo ambas las que tienen mayor cantidad tanto de admitidos, como de graduados. Después de estas dos áreas, siguen Matemáticas, Economía y Bellas Artes, en las siguientes posiciones, respectivamente. Por consiguiente, es claro que las especializaciones en Derecho siguen aportando una gran número de admitidos, pero que, en esta ocasión, los programas de Ingeniería presentaron un mayor número de personas que quisieron ingresar a la Universidad Nacional. Otra conclusión es que mientras para todas las otras áreas de conocimiento se mantuvo la proporción de personas admitidas con respecto a las que se gradúan, para Ingeniería podría significar que a pesar de que tiene un mayor número de admitidos, su proporción de admitidos/graduados no es tan buena, comparada con otras facultades.

Figura 4-4: Distribución de los admitidos a los programas de posgrado por su área de conocimiento



Ahora bien, una vez se tienen los datos de los admitidos completamente limpios, y con las mismas variables con las que se realizó el modelo de graduados, se procede a ejecutar el proceso para que con los 3 modelos ya obtenidos para los graduados (*Naive Bayes*, *Random Forest*, Árbol de decisión) se pueda asignar una etiqueta a cada uno de los nuevos registros, obteniendo así el diagrama de la Figura 4-5.

Figura 4-5: Diagrama en Orange para hacer la predicción de los admitidos

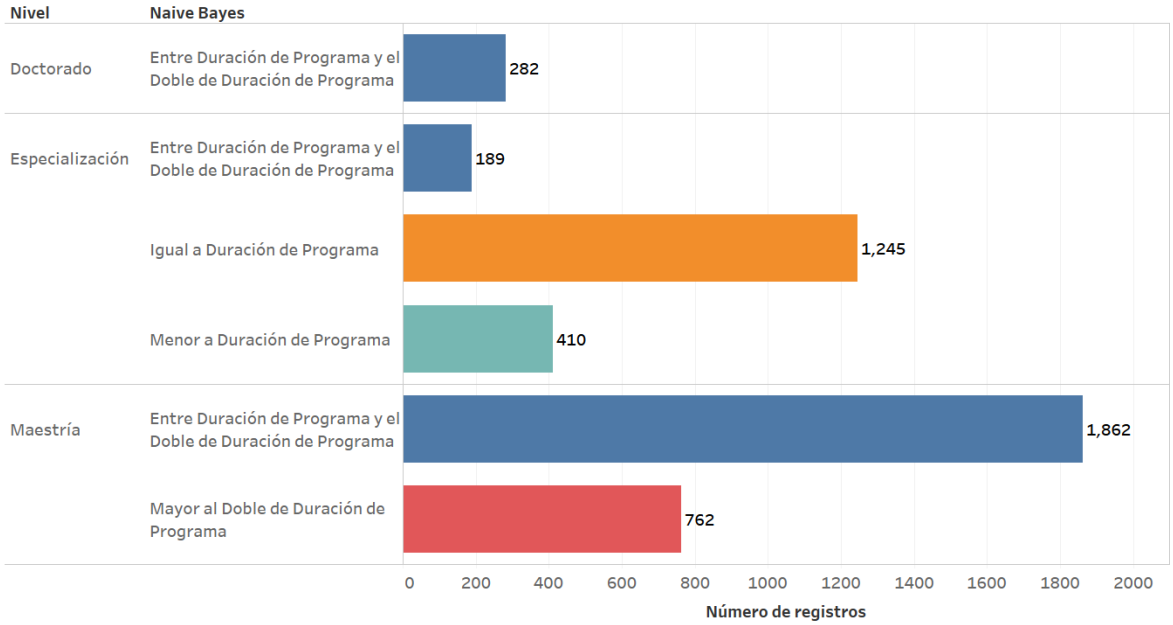


En este diagrama lo que se hace es cargar la información de los admitidos, después hacer una reducción tanto de columnas y variables, que tampoco se aplicaron en el modelo de los graduados, después se aplica la función de predicción, la cual tiene como entrada el modelo con los tres algoritmos utilizados y los datos a predecir, y como salida da el mismo conjunto de datos, con la etiqueta que le asigna cada uno de los algoritmos y con la probabilidad que le cada algoritmo a cada una de las clases. Una vez se obtiene esta predicción, y como se observa en la Figura 4-5, se procede a organizar todos los datos, con las etiquetas y las probabilidades dadas por cada algoritmo para así poder producir un archivo de salida; esto con el fin de que el listado se pueda consultar por los programas en el Sistema de Autoevaluación de una manera sencilla.

Al analizar las predicciones hechas por los tres algoritmos, y teniendo en cuenta que el que da los resultado más confiables fue el Random Forest, se puede observar cómo difieren bastantes en sus predicciones. Aunque esta situación era de esperar teniendo en cuenta los resultados de la evaluación del modelo. En la Figura 4-6, por ejemplo, se pueden observar los resultados de la predicción hecha por el algoritmo Naive Bayes. En este gráfico se ve por ejemplo que la totalidad de los admitidos a programas de Doctorado fueron clasificados en una sola clase, y que los otros niveles de posgrado si fueron separados en las diferentes clases; eso sí, este algoritmo tuvo coherencia para los niveles de Especialización y Maestría, ya que viendo los datos de los graduados,

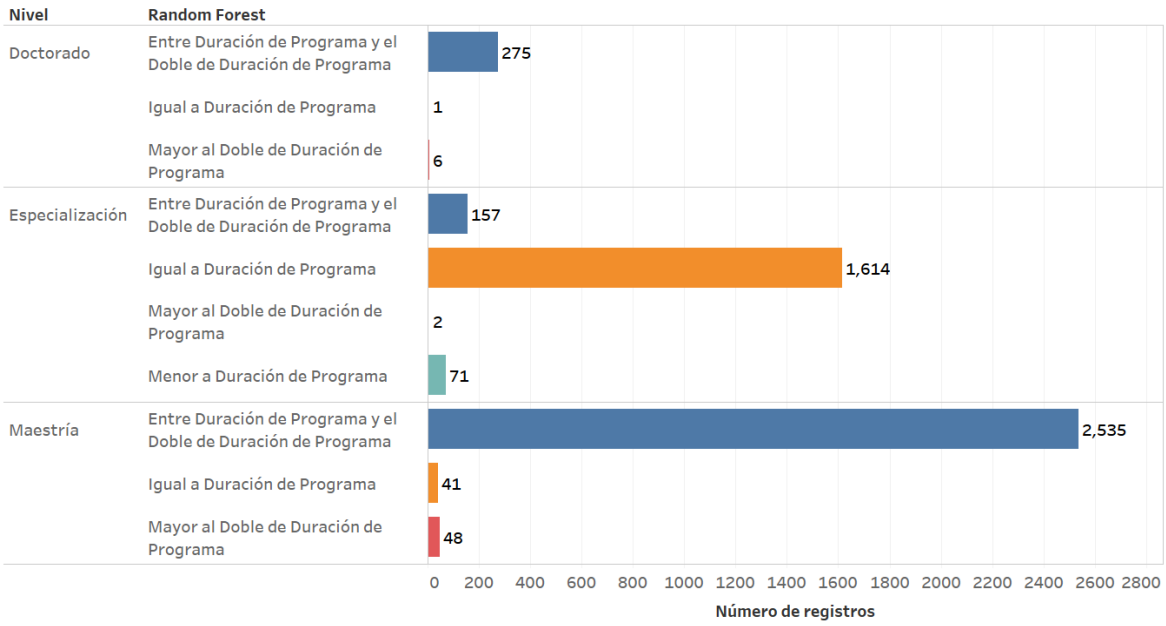
la mayoría de los de Especialización tenían etiqueta “Igual a Duración de Programa”, y, de igual manera, la clase “Entre Duración de Programa y doble de Duración de Programa” es la que predominaba para los programas de Maestría.

Figura 4-6: Predicción de los admitidos por nivel de programa (*Naive Bayes*)



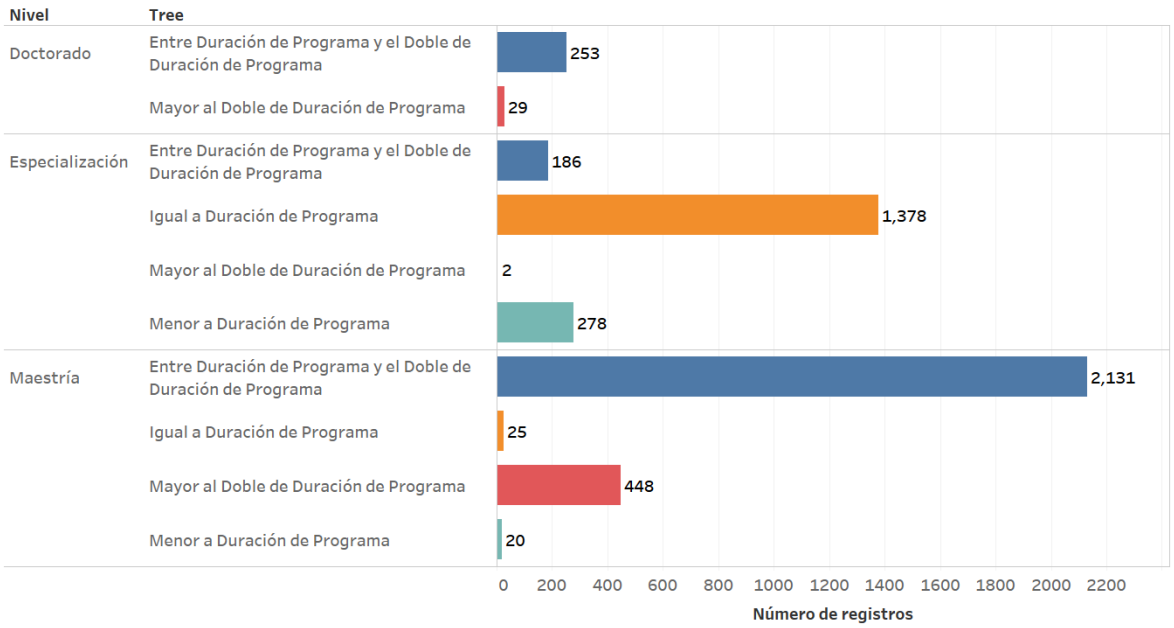
En la Figura 4-7 se observa la predicción que se realizó con ayuda del algoritmo *Random Forest,* al ser el algoritmo que mejor resultados obtuvo con las medidas de rendimiento y curvas ROC, es el algoritmo más confiable para poder hacer este tipo de predicciones. Aquí también se observa la tendencia de las clases “Igual a Duración de Programa” y “Entre duración de programa y el doble de duración de programa" para Maestría y Especializaciones respectivamente; aunque, en este caso, esas clases presentaron bastante diferencia al compararlas con las otras clases. Para los doctorados, este algoritmo también clasificó la mayoría de los admitidos en la clase “Entre Duración de Programa y doble de Duración de Programa”, pero se tuvieron 6 casos que los clasificó en “Mayor al Doble de Duración de Programa”, luego los programas deben prestar atención a ese tipo de casos, para prevenirlos.

Figura 4-7: Predicción de los admitidos por nivel de programa (*Random Forest*)



Finalmente, se puede observar cómo el algoritmo de árbol de decisión (ver Figura 4-8), a pesar de tener resultados de evaluación cercanos al *Random Forest,* en esta predicción si difiere un poco con respecto a lo que plantea el algoritmo anterior. Por ejemplo, la cantidad de personas clasificadas en otra clase que no fuera la predominante en cada nivel de estudios, sube y con gran diferencia, ya que tanto para especializaciones como para maestrías, con el algoritmo *Random Forest* esa cifra no pasa de 100 casos; en cambio, el algoritmo de árbol de decisión pasa esa cifra solo teniendo en cuenta una clase. En maestrías, por ejemplo, el número de personas que el algoritmo predice como graduados en un tiempo mayor al doble de la duración del programa es de 448, difiriendo bastante con la predicción de *Random Forest* de 48, y muy baja con respecto a la que da *Naive Bayes* de 762 personas.

Figura 4-8: Predicción de los admitidos por nivel de programa (Árbol de Decisión)



En conclusión y siguiendo la lógica de los resultados de las medidas de evaluación y las curvas ROC, se decidió mostrar solamente los resultados de la predicción hecha por el algoritmo *Random Forest*, para que los programas pueda visualizar la predicción que se le realizó a los estudiantes recién admitidos.

4.2 Módulo en el Sistema de Autoevaluación

Actualmente, la Dirección Nacional de Programas de Posgrado cuenta con el Sistema de Autoevaluación para Programas de Posgrado (SAPUN), en el cual se agrupa toda la información y documentos necesarios para que cada programa pueda llevar a cabo un proceso de auto-evaluación o acreditación, según sea el caso de cada programa. A este sistema solo tienen acceso algunas dependencias, como la propia Dirección de Posgrados, la Dirección de Planeación Nacional, y la Vicerrectoría Académica y, en cada sede, las Direcciones Académicas también tienen acceso al sistema. En cuanto a facultades, las Vicedecanaturas tienen acceso, los profesores directores de departamento y los coordinadores de cada programa, que están entre los usuarios más importantes del Sistema, ya que son ellos los que pueden ingresar tanto la ponderación y

clasificación de los diferentes factores y características que forman parte del modelo de autoevaluación, y, además, son los que registran y definen sus planes de mejoramiento.

Figura 4-9: Módulos principales del Sistema de Autoevaluación de Posgrados

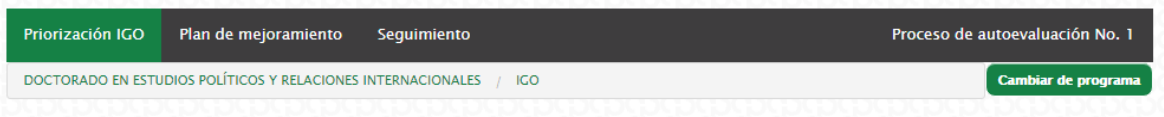
Tomado de SAPUN

En la Figura 4-9 se pueden visualizar los 5 módulos principales a los que puede tener acceso un usuario, en este caso un coordinador de programa. El módulo de recolección permite ingresar información de diferente índole perteneciente a cada programa y que no está disponible en las fuentes oficiales, como los diferentes proyectos y trabajos de investigación que se han realizado, los semestres en los que los profesores han estado vinculados al programa, entre otros datos. El módulo de Indicadores es aquel en el cual se pueden visualizar todos los indicadores estipulados en el modelo de autoevaluación y cuya fuente puede venir o de fuentes oficiales o de la información registrada por cada programa en el módulo anterior. El módulo de ponderación es en el que los programas pueden dar una ponderación y una calificación a cada uno de los factores y características dándole así un puntaje general de importancia a cada uno de ellos. El módulo de informe final es el que le permite a los programas registrar textos de introducción y conclusiones,

así como mostrar los resúmenes de los diferentes datos y bases que tiene el sistema en los llamados Cuadros Maestros. El objetivo de este módulo es generar el informe final para cada programa.

El cuarto módulo es el de Plan de Mejoramiento (ver Figura 4-10), en el los programas pueden registrar el plan que fue diseñado por ellos mismos, tener una priorización IGO, hacer seguimiento de las acciones y empezar a registrar nuevas acciones que permitan el total cumplimiento del plan. Se hace énfasis en este módulo ya que en él se abrió una nueva pestaña (ver Figura 4-11) que los programas pueden consultar para poder revisar los resultados de esta investigación.

Figura 4-10: Opciones disponibles en el módulo de Plan de Mejoramiento



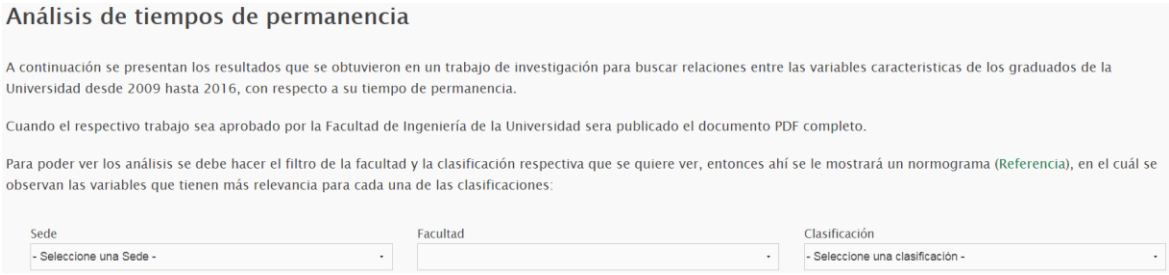
Tomado de SAPUN

Figura 4-11: Visualizaciones disponibles en la opción “Plan de mejoramiento”

Tomado de SAPUN

Para visualizar los resultados de esta investigación, en primer, lugar se presenta (ver Figura 4-12) una pequeña introducción de lo que se hizo en la investigación, haciendo referencia a lapso de tiempo que fue analizado y los resultados que se mostrarán posteriormente. También, se planea publicar este documento en esa ventana para que los programas puedan tener más referencia del trabajo realizado y, así mismo, puedan entender los resultados.

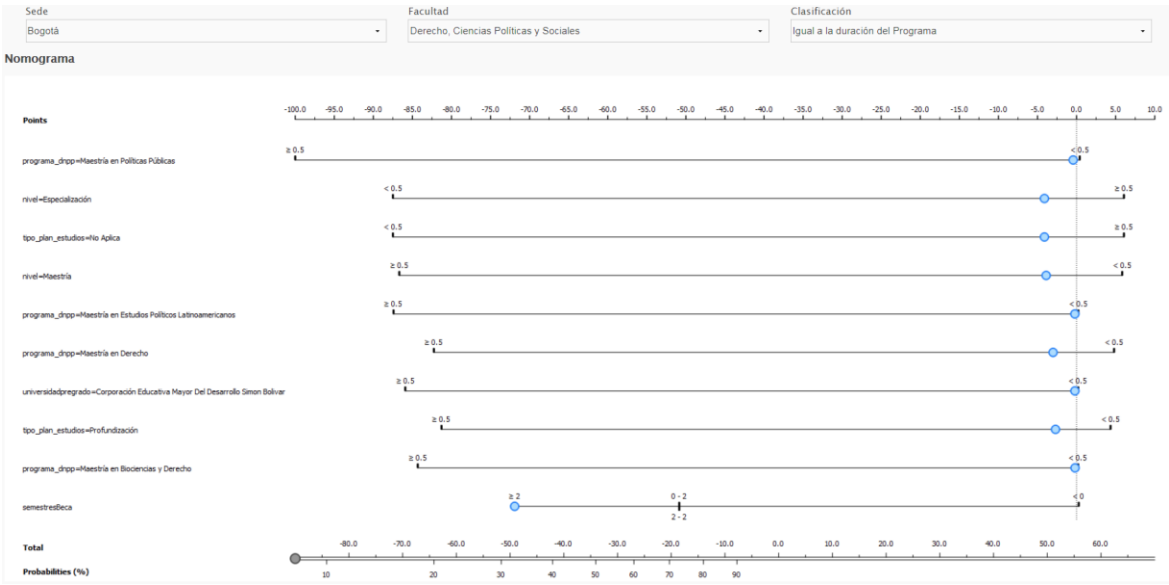
Figura 4-12: Visualización de la opción análisis de permanencia



Tomado de SAPUN

Una vez pasa el texto introductorio, el programa se encuentra con un pequeño filtro para que cada uno pueda visualizar el nomograma que sea propio de cada Facultad y cada Clase (ver Figura 4-13). A pesar de que según los resultados el nomograma, que es la visualización del resultado de aplicar *Naive Bayes,* no presenta los resultados del mejor algoritmo de clasificación, se decidió presentarlo a los programas, debido a que las medidas de rendimiento de los 3 algoritmos son muy cercanas luego debería tener resultados parecidos a los que pueden obtener los 2 algoritmos basados en árboles de decisión; además, su forma de visualizar los resultados hacen que se pueda presentar de una manera más sencilla y fácil de leer a los programas, a diferencia de lo que pueden ser los grandes y anchos árboles de decisión.

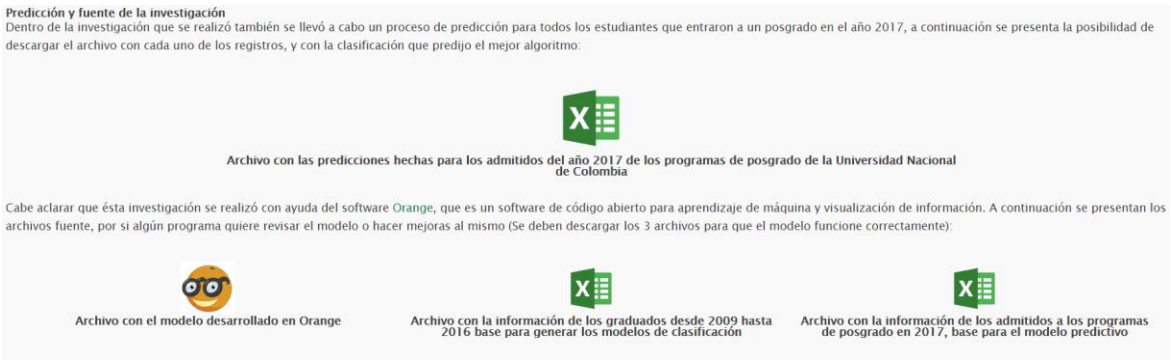
Figura 4-13: Presentación de Nomogramas en el SAPUN para cada Facultad



Tomado de SAPUN

Después de presentar la sección de nomograma, que da cuenta de la visualización de los resultados del modelo obtenido para los datos de graduación, se presenta una sección de descargas (ver Figura 4-14), en la cual, en primer lugar, se deja disponible la predicción hecha a los estudiantes admitidos en este 2017, para los programas de posgrado de la Universidad, con cada uno de los registros incluyendo la etiqueta dada por el *Random Forest*, y con la probabilidad dada por el mismo algoritmo a cada una de las clases. También, se dejan disponibles los tres insumos básicos de este trabajo: el modelo total hecho en Orange, la base de graduados desde 2009 hasta 2016 utilizada para generar el modelo de clasificación, y la base de admitidos de 2017, que es el archivo base para poner a prueba los tres algoritmos de clasificación y tratar de predecir cuál podría ser el futuro de estos nuevos integrantes de los programas de posgrado de la Universidad Nacional de Colombia.

Figura 4-14: Presentación de los datos de predicción y los archivos fuente de la investigación



Tomado de SAPUN

5. Conclusiones y recomendaciones

5.1 Conclusiones

En el mundo actual, en donde se obtienen datos en gran cantidad cada segundo, hacer análisis y obtener información adicional o conocimiento de los datos puede marcar la diferencia no solo para obtener ventajas competitivas, sino también para que a nivel interno de una organización pueda mejorar el proceso de la toma de decisiones. Cuando ese descubrimiento de conocimiento se aplica al ámbito educativo, los objetivos no son muy diferentes, y se trata de caracterizar los comportamientos de los estudiantes, mejorar los procesos de aprendizaje, mejorar indicadores de graduación, bajar tasas de deserción, entre otras metas posibles.

La aplicación de la minería de datos educacional a un caso específico como el de los graduados de programas de posgrado de la Universidad Nacional de Colombia, entre otras cosas, tiene como objetivo dar a conocer a los mismos programas características patrones que se pueden obtener de las características académicas y no académicas del conjunto de datos, y que les permita a los tomadores de decisiones generar estrategias académicas y de bienestar para mejorar el tiempo de permanencia de los actuales estudiantes. En la búsqueda de patrones, hay variables que inciden mucho más que otras, el nivel del programa parece ser una de ellas, ya que los estudiantes de especialización en un alto porcentaje terminan sus programas en el tiempo de duración reglamentado por el programa, o en ocasiones lo terminan, antes del tiempo previsto. Por ejemplo, se encontró que las mujeres que ingresan a una especialización son de los casos en los cuales terminar el programa antes del tiempo reglamentado por el programa presenta una opción muy probable. Algo similar pasa con las personas que al momento de iniciar una especialización son solteros o que se hayan graduado de la Universidad Nacional. Por el contrario, los estudiantes que ingresan a las maestrías bajo el plan de estudios de investigación tienen una mayor probabilidad

de graduarse en el más del doble de la duración de su programa, y si a esas condiciones se le suma que los estudiantes son hombres, la posibilidad aumenta.

Otro patrón encontrado es que en un 79% de los casos, un estudiante que estudie alguna maestría del área de conocimiento de Matemáticas y Ciencias Naturales se demora más de lo que el programa tiene reglamentado para su duración. Y un caso al que hay que prestar gran atención son los estudiantes que vienen de un programa de pregrado que tenga que ver con Educación, ya que en el 75% de los casos, tampoco terminan su programa en el tiempo que debería ser, más específicamente si estudian el programa de Maestría en Enseñanza de las Ciencias Exactas y Naturales en cualquiera de las sedes que lo oferta, en el 82.8% de los casos su duración está comprendida entre la duración reglamentada por el programa y el doble del mismo.

En cuanto a la aplicación de minería de datos, se observó que existen varios algoritmos de clasificación, dentro de ellos están los que se usaron en este trabajo, Random Forest, Árbol de decisión y Naive Bayes, y se pudo concluir que para este conjunto de datos, aunque los 3 algoritmos tuvieron medidas de rendimiento cercanos, y comportamiento de las curvas ROC bastante similares, el algoritmo Random Forest si tiene una mejor rendimiento.

El proceso de predicción entonces, aunque se hizo con los tres algoritmos, se tomó la decisión de tomar en cuenta sólo las predicciones que hizo el algoritmo Random Forest con los resultados anteriormente descrito. Además, haciendo un análisis de las predicciones hechas por el algoritmo tiene gran similitud con los patrones que se obtuvieron en la clasificación hecha con los graduados, luego los resultados del algoritmo se toman como muy confiables.

5.2 Trabajo Futuro

Dentro del campo de la minería de datos educacional existen muchas aplicaciones, este trabajo puede servir como base para hacer un trabajo similar en otras universidades, o para dirigir la investigación a resolver nuevos problemas, por ejemplo, tratar de plasmar el mismo trabajo pero aplicado a los programas de pregrado, teniendo en cuenta las diferencias que se tienen entre los diferentes niveles de los programas.

Otra posibilidad de trabajo futuro se puede realizar con el uso de otros algoritmos de clasificación, como lo pueden ser un algoritmo de regresión lineal, redes neuronales, o máquinas de soporte vectorial.

Para mejorar este tipo de trabajos se podrían incluir otras variables, sobre todo de tipo no académico, como la situación laboral del graduado, poder obtener información de la situación familiar, u otro tipo de variables adicionales que puedan afectar la clasificación y las predicciones. Estas variables se podrían obtener teniendo en cuenta fuentes de información adicionales como el ICFES, o tener una comunicación directa con los graduados haciendo uso de encuestas, redes sociales u otros medios de comunicación.

Bibliografía

[1] "International Educational Data Mining Society." [Online]. Available: http://www.educationaldatamining.org/.

[2] O. Moscoso-zea and S. Luján-mora, "Minería de Datos Educacionales : una visión holística Educational Data Mining : an holistic view," in *Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on*, 2016.

[3] S. R. Timarán-Pereira, I. Hernández-Arteaga, S. J. Caicedo-Zambrano, A. Hidalgo-Troya, and C. J. AlvaradoPérez, "El proceso de descubrimiento de conocimiento en bases de datos," in *En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, Ediciones., no. 2016, Bogotá, 2016, pp. 63–86.

[4] J. M. Moine and D. A. S. Haedo, "Metodologías para el descubrimiento de conocimiento en bases de datos : un estudio comparativo," Universidad Nacional de La Plata, 2013.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[6] Q. L. Q. Luo, "Advancing Knowledge Discovery and Data Mining," *First Int. Work. Knowl. Discov. Data Min. (WKDD 2008)*, pp. 7–9, 2008.

[7] X. Liu, Z. Huang, and B. Tong, "Review on the data mining technology and the applications on financial analysis area," *2016 Int. Conf. Commun. Electron. Syst.*, pp. 1–7, 2016.

[8] D. M. Moreno Torres, "Sistema Predictivo para la detección de niñas y adolescentes con alto riesgo de quedar en embarazo," 2016.

[9] B. E. V Comendador, L. W. Rabago, and B. T. Tanguilig, "An educational model based

on Knowledge Discovery in Databases (KDD) to predict learner's behavior using classification techniques," in *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2016, pp. 1–6.

[10] M. Ramzan and M. Ahmad, "Evolution of Data Mining : An Overview Evolution of Data Magnitude," *IT Business, Ind. Gov.*, 2014.

[11] A. N. Dharsandiya and M. R. Patel, "A review on Frequent Itemset Mining algorithms in social network data," *2016 Int. Conf. Wirel. Commun. Signal Process. Netw.*, pp. 1046–1048, 2016.

[12] M. A. Jayanthi, "Research Contemplate on Educational Data Mining," pp. 110–114, 2016.

[13] R. Marital and I. Cheat, "Data Mining : Data Lecture Notes for Chapter 2 Introduction to Data Mining by What is Data ?," *Ratio*, 2004.

[14] M. Kantardzic, "Data-Mining Concepts," *Data Min.*, pp. 1–25, 2011.

[15] E. León Guzmán, "Preprocesamiento de Datos," *Univ. Nac. Colomb.*, 2016.

[16] Q. Yu, H. Jiang, and Y. Yu, "Classification mining in securities investment cost," *Proc. - 3rd Int. Conf. Multimed. Inf. Netw. Secur. MINES 2011*, no. 2, pp. 205–209, 2011.

[17] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Predicting phishing websites using classification mining techniques with experimental case studies," *ITNG2010 - 7th Int. Conf. Inf. Technol. New Gener.*, pp. 176–181, 2010.

[18] M. Wang, B. Iyer, and J. S. Vitter, "Scalable Mining for Classification Rules in Relational Databases," *Lect. Notes-Monograph Ser.*, vol. 45, pp. 348–377, 2004.

[19] W. Wu and M. Peng, "A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-Term Wind Power Forecasting," *IEEE Internet Things J.*, vol. 4, no. 4, pp. 979–986, 2017.

[20] U. F. Alias, N. B. Ahmad, and S. Hasan, "Mining of E-learning Behavior using SOM Clustering," no. 1998, pp. 1–4, 2017.

[21] F. Barrientos and S. Ríos, "Aplicación de minería de datos para predecir fuga de clientes en la industria de las telecomunicaciones," *Sist. Inf.*, vol. XXVII, p. 159, 2013.

[22] W. A. C. Rojas and C. M. Villegas, “Graphical representation and exploratory visualization for decision trees in the KDD process,” *Proc. - 2012 9th Electron. Robot. Automot. Mech. Conf. CERMA 2012*, pp. 203–210, 2012.

[23] S. H. Cha and S. N. Srihari, “A priori algorithm for sub-category classification analysis of handwriting,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2001–Janua, pp. 1022–1025, 2001.

[24] K. Goeschel, “Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis,” in *Conference Proceedings - IEEE SOUTHEASTCON*, 2016.

[25] P.-N. Tan, M. Steinbach, and V. Kumar, “Association Analysis: Basic Concepts and Algorithms,” *Introd. to Data Min.*, pp. 327–414, 2005.

[26] C. E. Lopez Guarin, E. L. Guzman, and F. A. Gonzalez, “A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining,” *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 10, no. 3, pp. 119–125, 2015.

[27] P.-N. Tan, M. Steinbach, and V. Kumar, “Classification: Basic Concepts, Decision Trees, and Model Evaluation,” in *Introduction to Data Mining*, vol. 67, no. 17, 2006, pp. 145–205.

[28] S. Moedjiono, Y. R. Isak, and A. Kusdaryono, “Customer loyalty prediction in multimedia Service Provider Company with K-Means segmentation and C4.5 algorithm,” *2016 Int. Conf. Informatics Comput. ICIC 2016*, no. Icic, pp. 210–215, 2017.

[29] Akella, AlanUS, AllenDowney, A. Lincoln, Arauzo, Arichnad, Bewildebeast, and Bkell, “Naive Bayes classifier,” pp. 1–9, 2006.

[30] M. Karim and R. M. Rahman, “Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing,” *J. Softw. Eng. Appl.*, vol. 6, no. 4, pp. 196–206, 2013.

[31] S. Aygün and E. Olcay, “A Benchmarking : Feature Extraction and Classification of Agricultural Textures Using LBP , GLCM , RBO , Neural Networks , k-NN , and Random

Forest," *6th Int. Conf. Agro-Geoinformatics*, 2017.

[32] M. Tirados, "Algoritmo Random Forest," *Big Data Hispano*, 2014. [Online]. Available: http://www.bigdatahispano.org/noticias/algoritmo-random-forest/.

[33] K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in *ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology*, 2015.

[34] Á. Jiménez Galindo and H. Álvarez García, "Minería de Datos en la Educación," *Intel. en Redes Comun.*, 2010.

[35] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, vol. 3536, no. c, pp. 1–1, 2017.

[36] T.-S. Xu, H.-D. Chiang, G.-Y. Liu, and C.-W. Tan, "Hierarchical K-means Method for Clustering Large-Scale Advanced Metering Infrastructure Data," *IEEE Trans. Power Deliv.*, vol. 32, no. 2, pp. 609–616, 2017.

[37] E. Hot and V. Popovi, "Soil data clustering by using K-means and fuzzy K-means algorithm," *Telecommun. Forum Telfor (TELFOR), 2015 23rd*, pp. 890–893, 2015.

[38] A. F. Elgamal, "An Educational Data Mining Model for Predicting Student Performance in Programming Course," *Int. J. Comput. Appl.*, vol. 70, no. 17, pp. 22–28, 2013.

[39] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," *Commun. Netw. Technol. (ICCNT), 2014 Int. Conf. Comput. Intell. Comput. Res.*, pp. 113–118, 2014.

[40] M. S. Halawa, "Predicting Student Personality Based on a Data- Driven Model from Student Behavior on LMS and Social Networks," pp. 294–299, 2015.

[41] O. N. Pratiwi, "Predicting student placement class using data mining," *Proc. 2013 IEEE Int. Conf. Teaching, Assess. Learn. Eng. TALE 2013*, no. August, pp. 618–621, 2013.

[42] A. González-Nucamendi, J. Noguez, L. Neri, and V. Robleda-Rella, "Predictive models to enhance learning based on student profiles derived from cognitive and social

constructs," in *Proceedings of 2015 International Conference on Interactive Collaborative and Blended Learning, ICBL 2015*, 2016, pp. 5–12.

[43] P. D. Schalk, D. P. Wick, P. R. Turner, and M. W. Ramsdell, "Predictive assessment of student performance for early strategic guidance," in *Proceedings - Frontiers in Education Conference, FIE*, 2011.

[44] L. Talavera and E. Gaudioso, "Mining student data to characterize similar behavior groups in unstructured collaboration spaces," in *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence*, 2004, no. June, pp. 17–23.

[45] A. Rodriguez, R. Pérez, J. Carrascal, and D. Durán, *Permanencia Estudiantil en los Postgrados de la Universidad Nacional de Colombia*. 2010.